

Medicina personalizada en la práctica clínica: manejo del big data genómico

Máster Universitario en Biotecnología Biomédica UPV
Valencia 17 y 21 Dic 2015



PRINCIPE FELIPE
CENTRO DE INVESTIGACION

Computational • Genomics



Príncipe Felipe Research Center

Goal: biomedical research

- **Basic research** in genes, targets, molecular and cellular processes, Nanomedicine and Computational Medicine
- **Translation into clinical practice:** personalized medicine, cancer, rare diseases, metabolic and functional impairment

<http://www.cipf.es/>

Who are we?

- The **Computational Genomics** Department, in Research Center Prince Felipe
- **Team:** multidisciplinary group of 14 researchers and technicians led by Joaquín Dopazo

<http://bioinfo.cipf.es/>

Who are we?



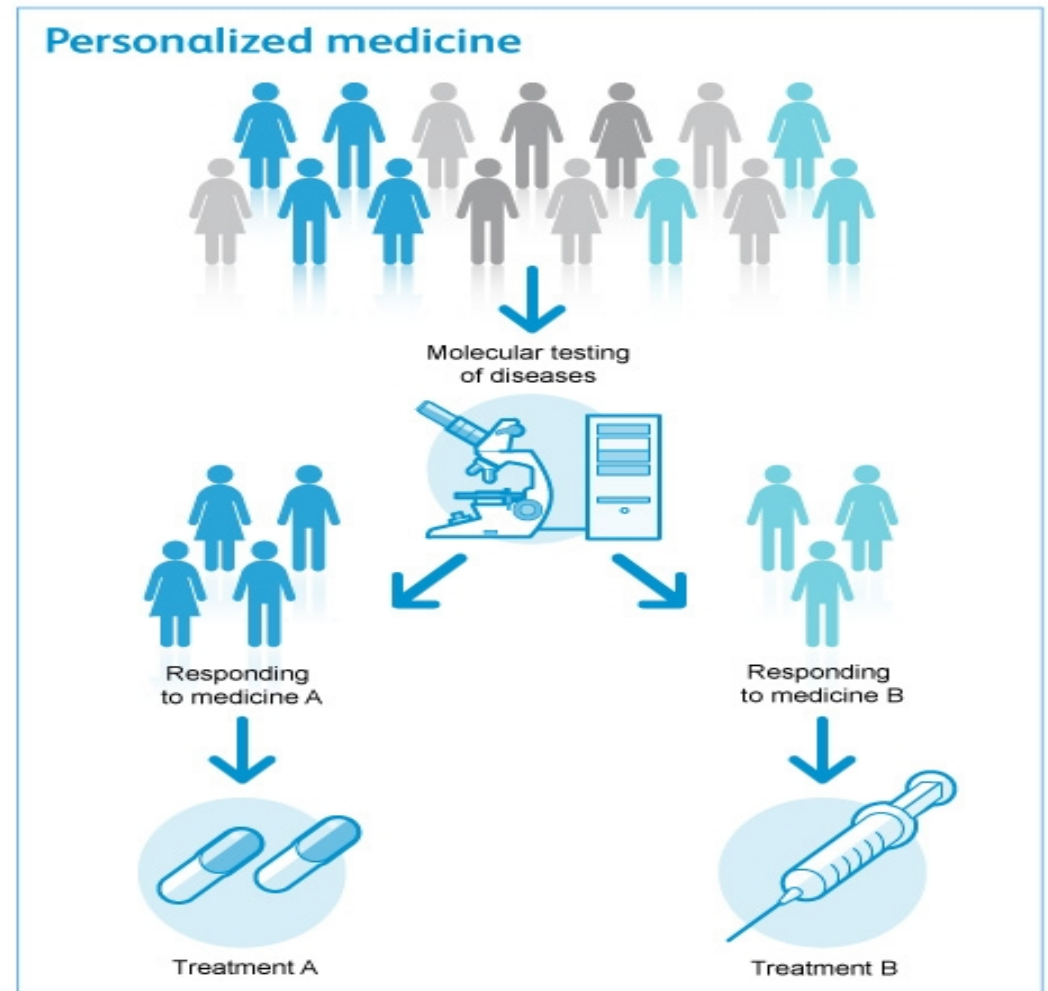
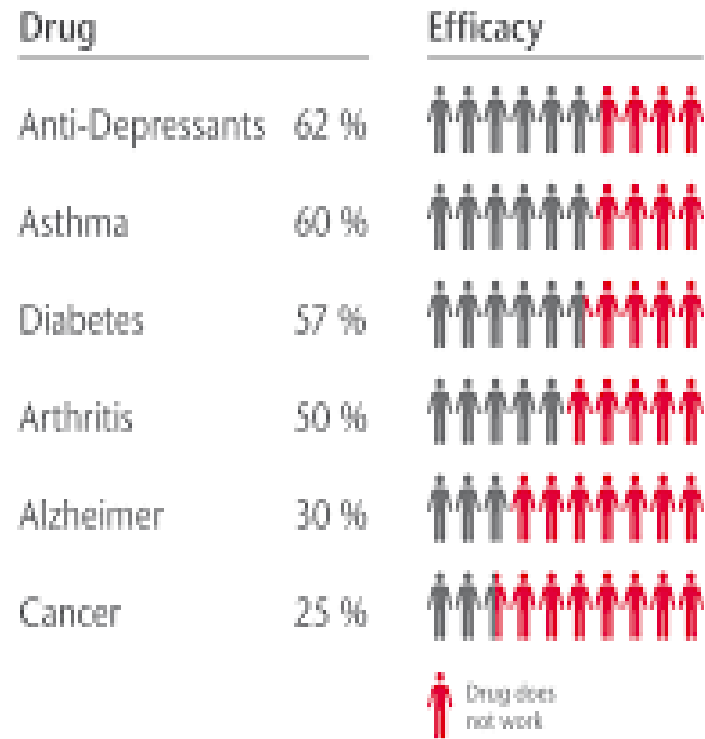
Introduction

Genomic Computational Department

Why are we interested in Computational Genomics?

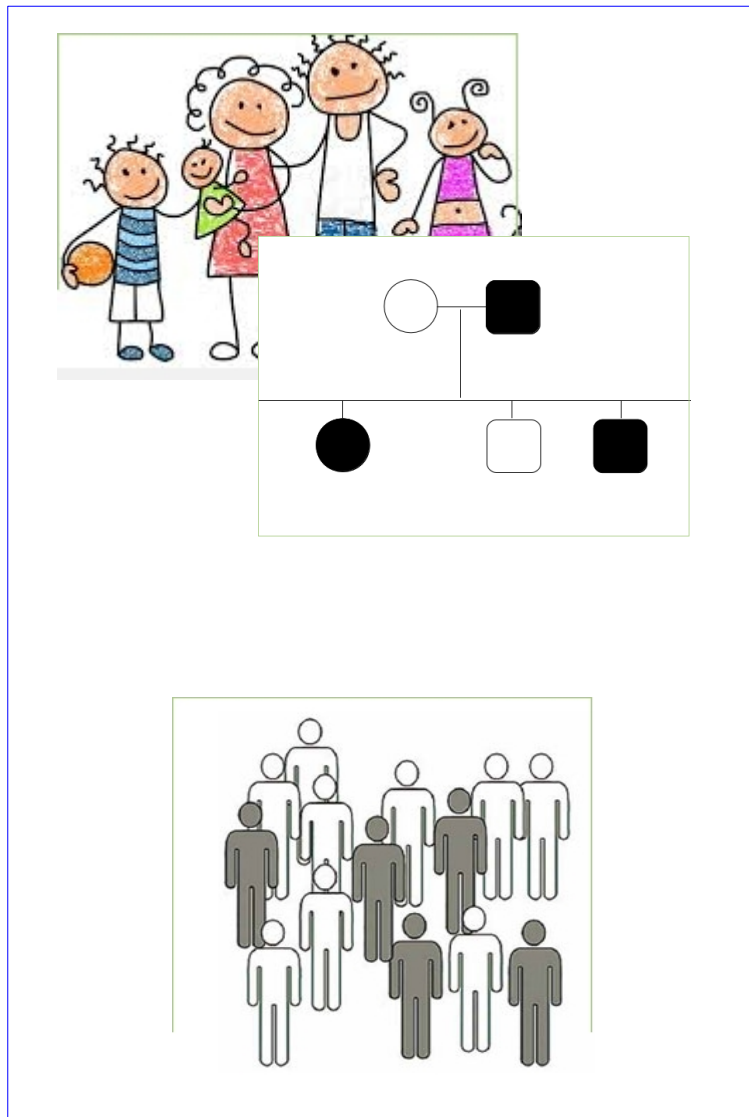
- The **overall goal** of the department:
 - ▢ Apply computational methods to biomedical and biotechnological problems
- **Research interests:**
 - ▢ The development and application of novel bioinformatics **methods** aimed at **discovering new drugs**
 - ▢ Identification of genes or proteins may be considered **therapeutic targets**
 - ▢ **Personalized medicine:** tools for discovering and diagnostic

Why are we interested in Computational Genomics?



New molecular and diagnostic technologies can be used to match select groups of patients with treatments that may give them the best results

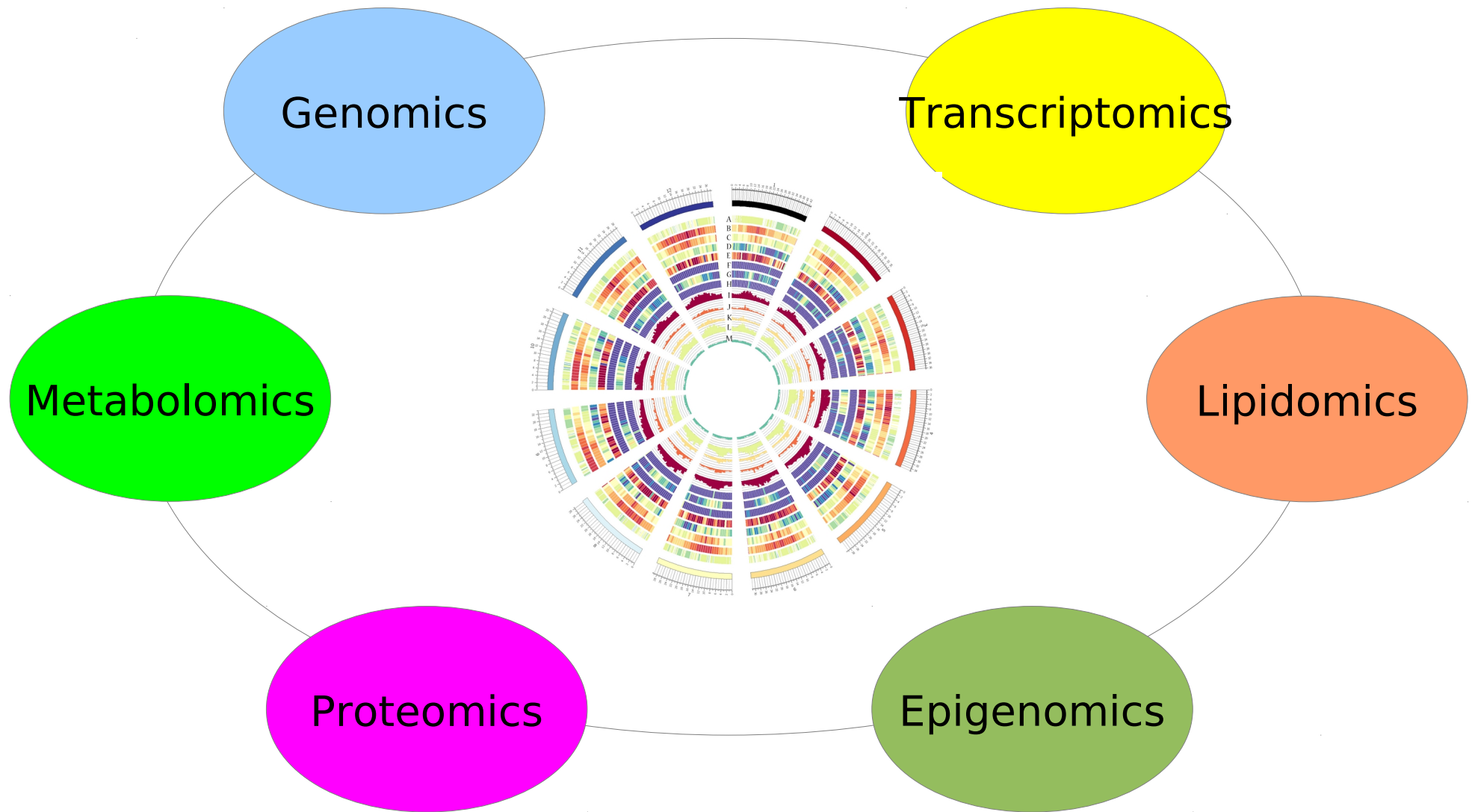
Why are we interested in Computational Genomics?



Introduction

Personalized Medicine and Mendelian Diseases

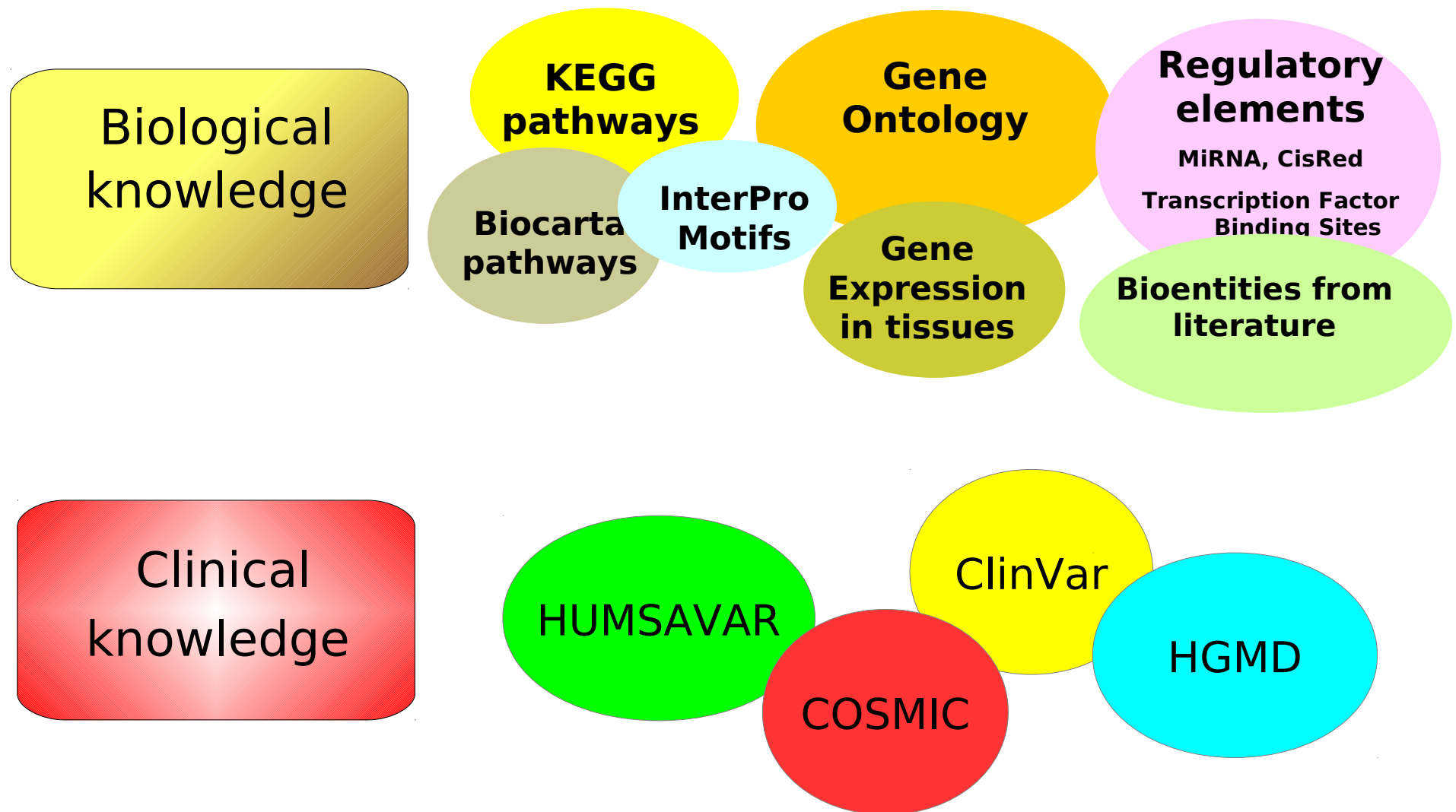
Big Data



Introduction

Omics sciences

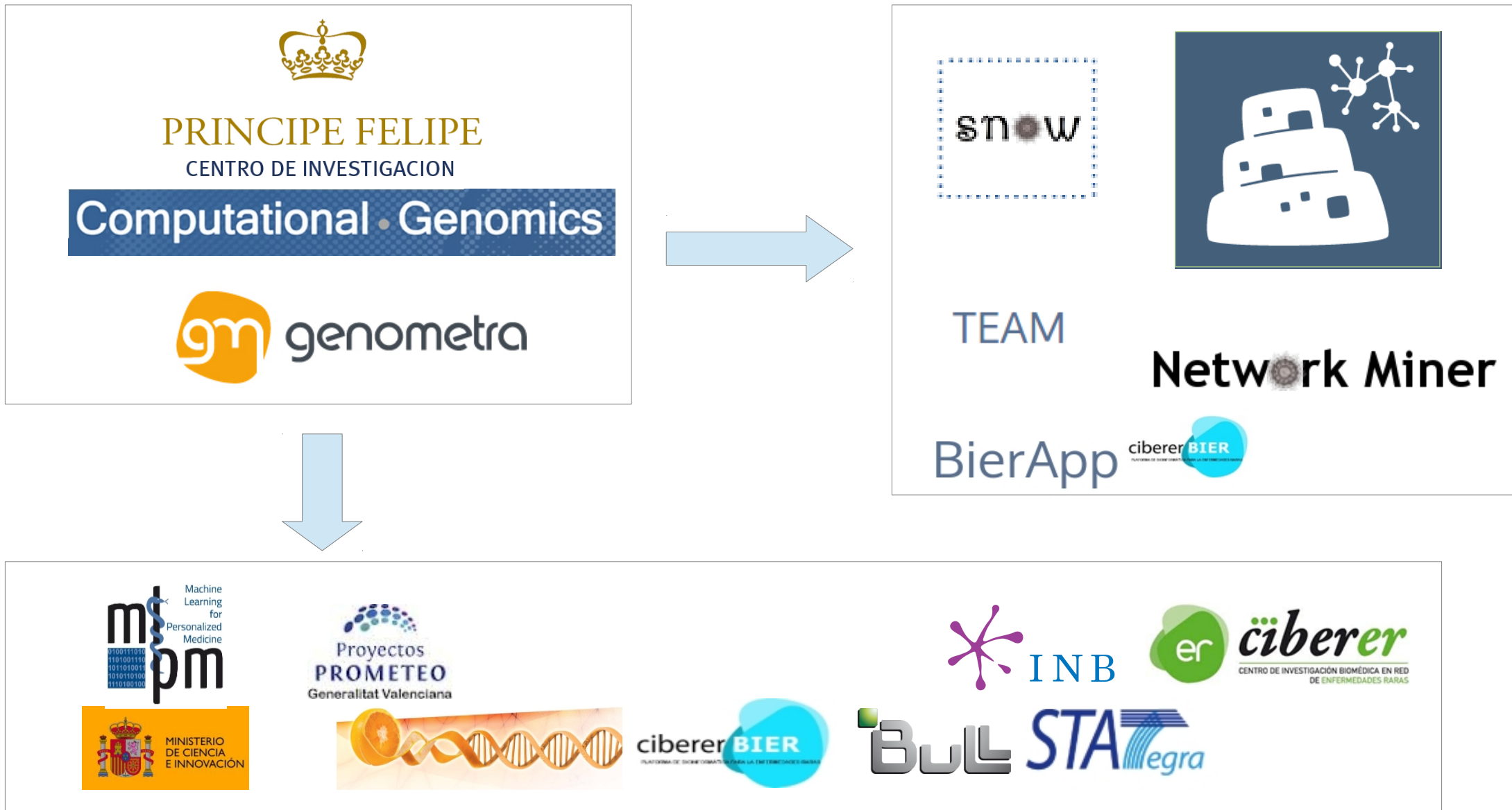
Big Data



How do we work?

- Our department collaborates in different research projects and converts customer needs into bioinformatics solutions
- Free software for several reasons:
 - **Any customer can try our tools**
 - **The scientific community can test our software**
 - This is the **current trend in Computational Genomics**

How do we work?



Introduction

Genomic Computational Department

How do we work?

IT4Innovations#
národní01\$!@%#0
superpočítačové
centrum\$@0@\$0%\$



OMICS MASTER

How do we work?

- El CIBER en su Área Temática de Enfermedades Raras (CIBERER) es el **centro de referencia** en España en investigación sobre **enfermedades raras**: <http://www.ciberer.es/>
- **Objetivo**: coordinar y favorecer la investigación básica, clínica y epidemiológica, así como potenciar que la investigación que se desarrolla en los laboratorios llegue al paciente, y dé respuestas científicas a las preguntas nacidas de la interacción entre médicos y enfermos.
- El CIBERER se compone de un equipo humano de más de 700 profesionales e integra a **62 grupos de investigación**.



How do we work?

- Curso CIBERER de análisis de datos genómicos,
28-30 Sep 2015 en Valencia:
<http://bioinfo.cipf.es/mda15ciberer>
- International course of Genomic Data Analysis,
29-4 Mar 2016, Valencia:
<http://bioinfo.cipf.es/gda16/program/>
- <http://bioinfo.cipf.es/courses>

Web tools to analyze omic data



Computational • Genomics



Web Tools for Genomic Data Analysis

1) Introduction to NGS Data Analysis

2) RNA-Seq Data Analysis

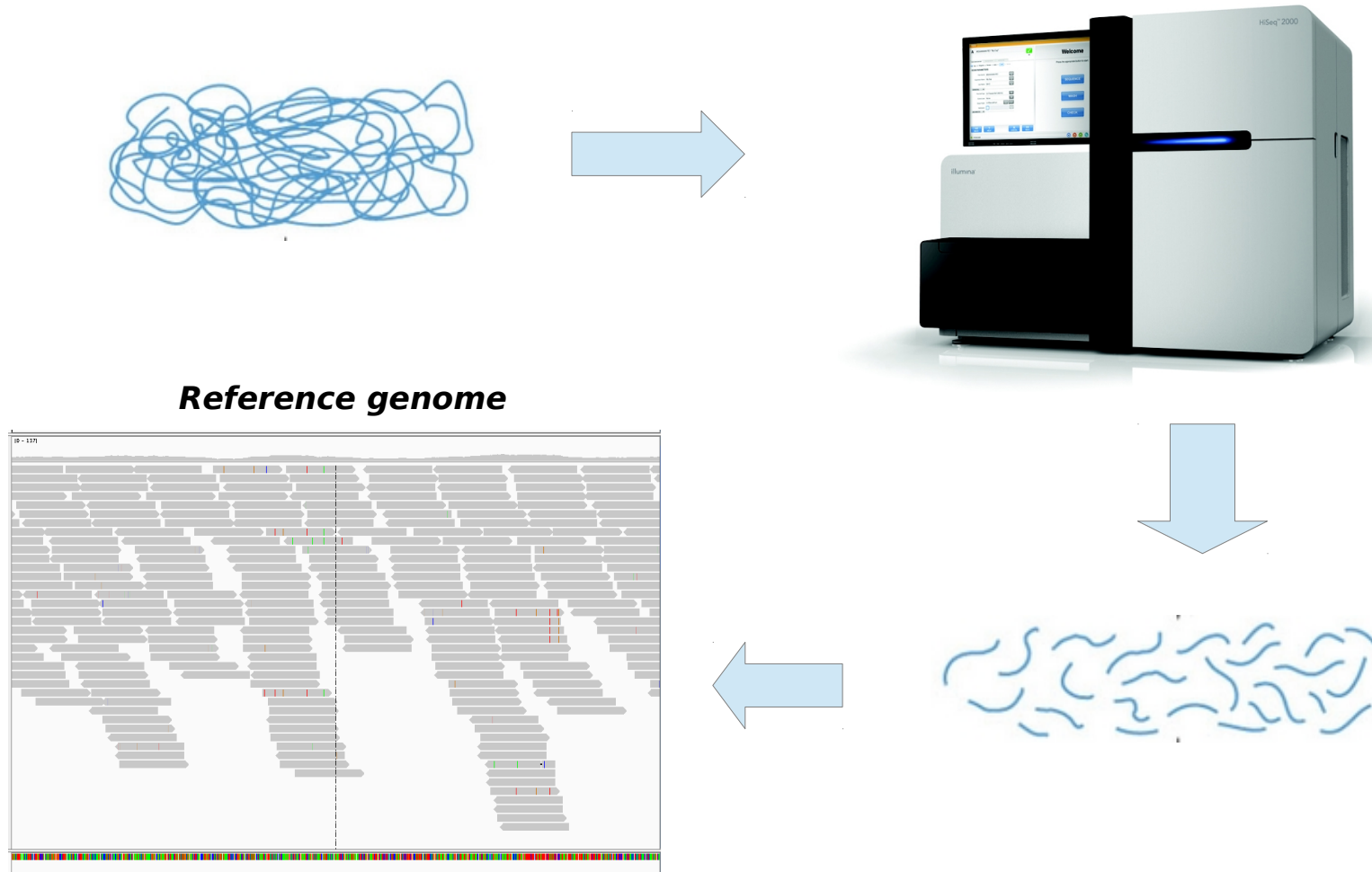
3) Resequencing Data Analysis

4) Omics Data Integration

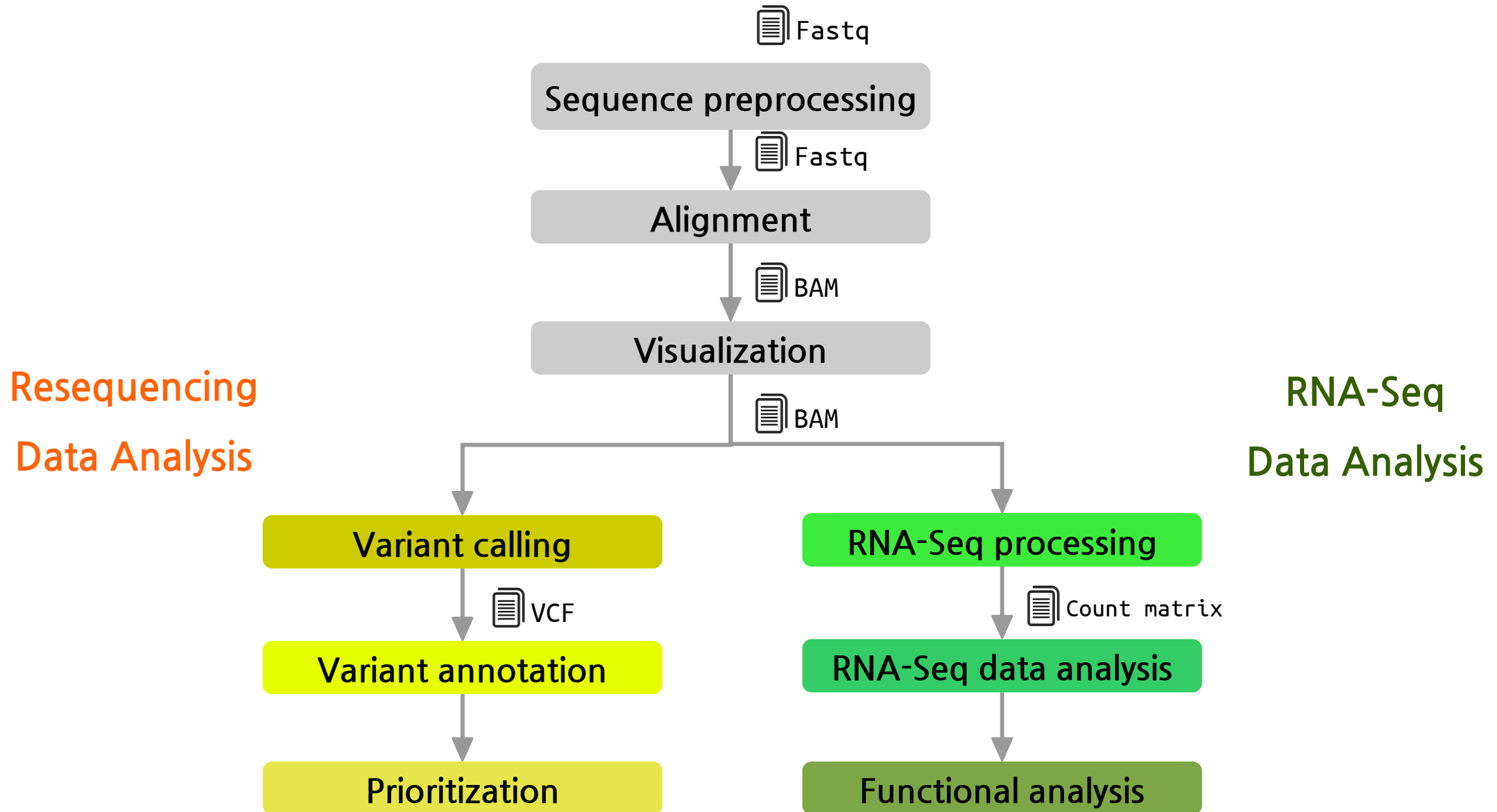
5) Functional Profiling

NGS technologies

How do these technologies work ?



NGS Data Analysis Pipeline



Fastq format

- We could say “it is a fasta with **qualities**”:
 - 1. Header (like the fasta but starting with “@”)
 - 2. Sequence (string of nt)
 - 3. “+” and sequence ID (optional)
 - 4. Encoded quality of the sequence

```
@SEQ_ID
GATTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTTGTTCAACTCACAGTTT
+
!''*(((((***+))%%%+))((%%%) .1***-+*'))**55CCF>>>>>CCCCCCC65
```

BAM/SAM format

```
@PG ID:HPG-Aligner VN:1.0
@SQ SN:20 LN:63025520

HWI-ST700660_138:2:2105:7292:79900#2@0/1 16 20 76703 254 76= * 0 0
GTTTAGATACTGAAAGGTACATACTTCTTTGTAGGAACAAGCTATCATGCTGCATTTCTATAATATCACATGAATA
GIJGJLGGFLILGGIEIFEKEDELIGLJIHJFIKKFELFIKLFFGLGHKKGJLFIIGKFFEFFEFGKCKFHHCCCF AS:i:254 NH:i:1 NM:i:0

HWI-ST700660_138:2:2208:6911:12246#2@0/1 16 20 76703 254 76= * 0 0
GTTTAGATACTGAAAGGTACATACTTCTTTGTAGGAACAAGCTATCATGCTGCATTTCTATAATATCACATGAATA
HHJFHLGFFLILEGIKIEEMGEDLIGLHIHJFIKKFELFIKLEFGKGHEKHJLFHIGKFFDFFEF GKDKFHHCCCF AS:i:254 NH:i:1 NM:i:0

HWI-ST700660_138:2:1201:2973:62218#2@0/1 0 20 76655 254 76M * 0 0
AACCCCAAAAATGTTGGAAGAATAATGTAGGACATTGCAGAAGACGATGTTTAGATACTGAAAGGGACATACTTCT
FEFFGHHHGGHFKCCJKFHIGIFFIFLDEJKGJGGFKIHLFIJGIEGFLDEDFLGEIIMHHIKL$BBGFFJIEHE AS:i:254 NH:i:1 NM:i:1

HWI-ST700660_138:2:1203:21395:164917#2@0/1 256 20 68253 254 4M1D72M * 0 0
NCACCCATGATAGACCAGTAAAGGTGACCACTTAAATTCCTTGCTGTGCAGTGTTCTGTATTCCTCAGGACACAGA
#4@ADEHFJFFEJDHJGKEFIHGHGBGFHHFIICEIIFFKIFHEGJEHHGLELEGKJMFGGGLEIKHLFGKIKHDG AS:i:254 NH:i:3 NM:i:1

HWI-ST700660_138:2:1105:16101:50526#6@0/1 16 20 126103 246 53M4D23M * 0 0
AAGAAGTGCAAACCTGAAGAGATGCATGTAAAGAATGGTTGGGCAATGTGCGGCAAAGGGACTGCTGTGTTCCAGC
FEHIGGHIGIGJI6FCFHJIFFLJJCJGJHGFKKKKGJIKHFFKIFFFKHFLKHGKJLJGKILLEFFLIHJIEIB AS:i:368 NH:i:1 NM:i:4
```

SAM Specification:

<http://samtools.sourceforge.net/SAM1.pdf>

Introduction

NGS data analysis: files format

VCF format

```
#fileformat=VCFv4.1
##fileDate=20090805
##source=myImputationProgramV3.1
##reference=file:///seq/references/1000GenomesPilot-NCBI36.fasta
##contig=<ID=20,length=62435964,assembly=B36,md5=f126cdf8a6e0c7f379d618ff66beb2da,species="Homo sapiens",taxonomy=x>
##phasing=partial
##INFO=<ID=NS,Number=1,Type=Integer,Description="Number of Samples With Data">
##INFO=<ID=DP,Number=1,Type=Integer,Description="Total Depth">
##INFO=<ID=AF,Number=A,Type=Float,Description="Allele Frequency">
##INFO=<ID=AA,Number=1,Type=String,Description="Ancestral Allele">
##INFO=<ID=DB,Number=0,Type=Flag,Description="dbSNP membership, build 129">
##INFO=<ID=H2,Number=0,Type=Flag,Description="HapMap2 membership">
##FILTER=<ID=q10,Description="Quality below 10">
##FILTER=<ID=s50,Description="Less than 50% of samples have data">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
##FORMAT=<ID=HQ,Number=2,Type=Integer,Description="Haplotype Quality">
#CHROM POS ID REF ALT QUAL FILTER INFO FORMAT NA00001 NA00002 NA00003
20 14370 rs6054257 G A 29 PASS NS=3;DP=14;AF=0.5;DB;H2 GT:GQ:DP:HQ 0|0:48:1:51,51 1|0:48:8:51,51 1/1:43:5:.,.
20 17330 . T A 3 q10 NS=3;DP=11;AF=0.017 GT:GQ:DP:HQ 0|0:49:3:58,50 0|1:3:5:65,3 0/0:41:3
20 1110696 rs6040355 A G,T 67 PASS NS=2;DP=10;AF=0.333,0.667;AA=T;DB GT:GQ:DP:HQ 1|2:21:6:23,27 2|1:2:0:18,2 2/2:35:4
20 1230237 . T . 47 PASS NS=3;DP=13;AA=T GT:GQ:DP:HQ 0|0:54:7:56,60 0|0:48:4:51,51 0/0:61:2
20 1234567 microsat1 GTC G,GTCT 50 PASS NS=3;DP=9;AA=G GT:GQ:DP 0/1:35:4 0/2:17:2 1/1:40:3
```

<http://www.1000genomes.org/>

Counts

Gene

Sample



Ensembl	Gene.Name	T1	T2	T3	T4	T5	WT1	WT2	WT3	WT4	WT5	WT6
ENSMUSG00000000134	Tfe3	312	295	333	258	392	257	344	223	423	277	389
ENSMUSG00000000142	Axin2	165	171	138	166	203	170	172	119	203	147	178
ENSMUSG00000000148	Brat1	213	196	207	224	350	204	268	143	300	177	288
ENSMUSG00000000149	Gna12	684	684	613	545	900	496	672	426	1023	583	797
ENSMUSG00000000154	Slc22a18	3	2	3	2	2	3	3	2	1	1	3
ENSMUSG00000000157	Itgb2l	0	0	0	0	0	0	0	0	0	0	0
ENSMUSG00000000159	Igsf5	0	0	0	0	0	0	0	0	0	0	0
ENSMUSG00000000167	Pih1d2	15	19	6	10	9	5	5	5	7	6	6
ENSMUSG00000000168	Dlat	899	777	967	756	1116	777	1047	614	1155	894	1126
ENSMUSG00000000171	Sdhd	1055	1003	1047	914	1430	939	1192	766	1390	916	1412
ENSMUSG00000000182	Fgf23	1	0	3	1	0	2	0	2	2	0	0
ENSMUSG00000000183	Fgf6	0	0	0	0	0	0	0	1	0	0	0
ENSMUSG00000000184	Ccnd2	1961	1978	1804	1779	2090	1655	2148	1585	2504	1895	2274
ENSMUSG00000000194	Gpr107	784	733	667	615	889	654	818	483	1034	627	1015
ENSMUSG00000000197	Nalcn	1120	1009	1047	917	1356	1129	1202	758	1625	1127	1044

Web Tools for Genomic Data Analysis

1) Introduction to NGS Data Analysis

2) RNA-Seq Data Analysis

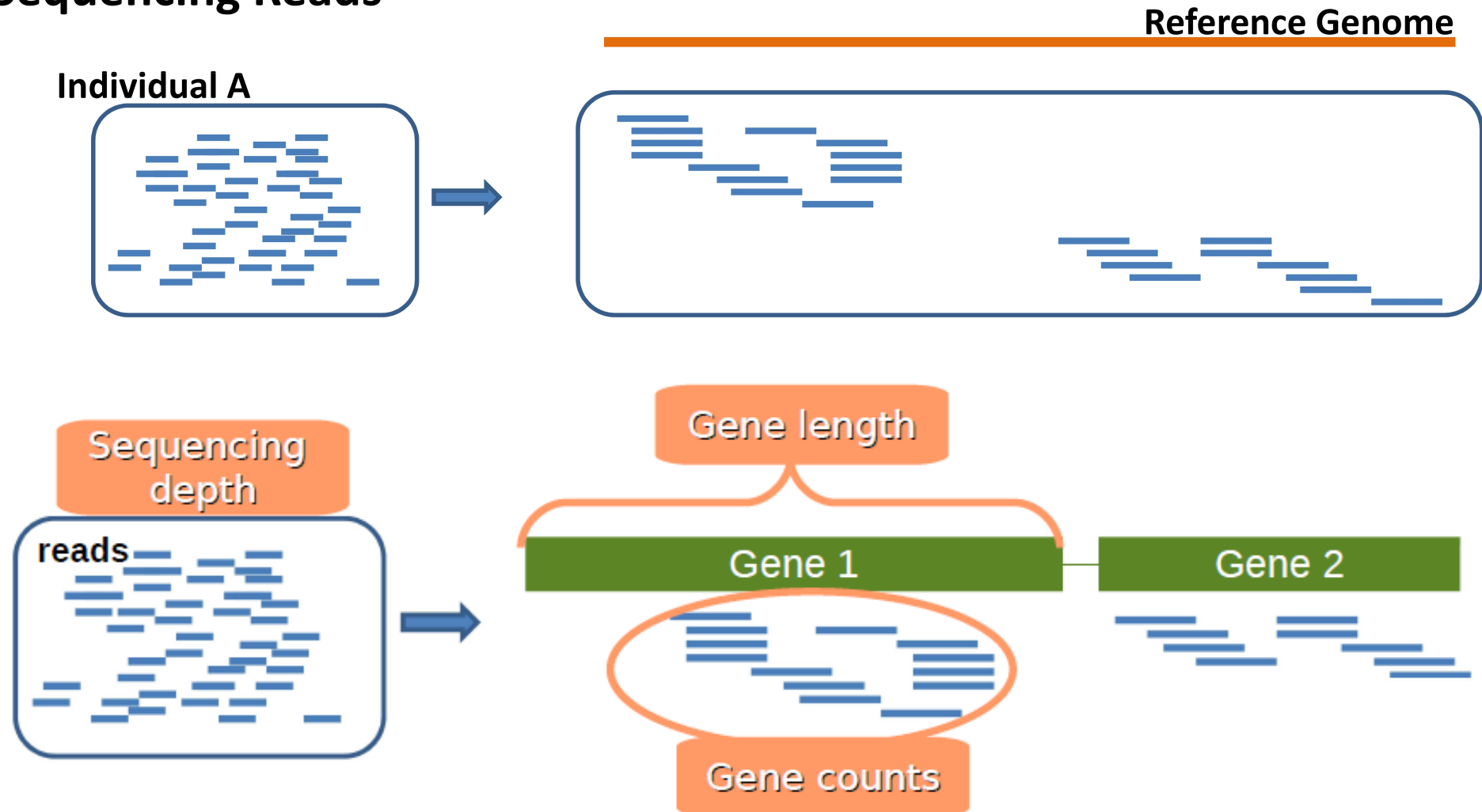
3) Resequencing Data Analysis

4) Omics Data Integration

5) Functional Profiling

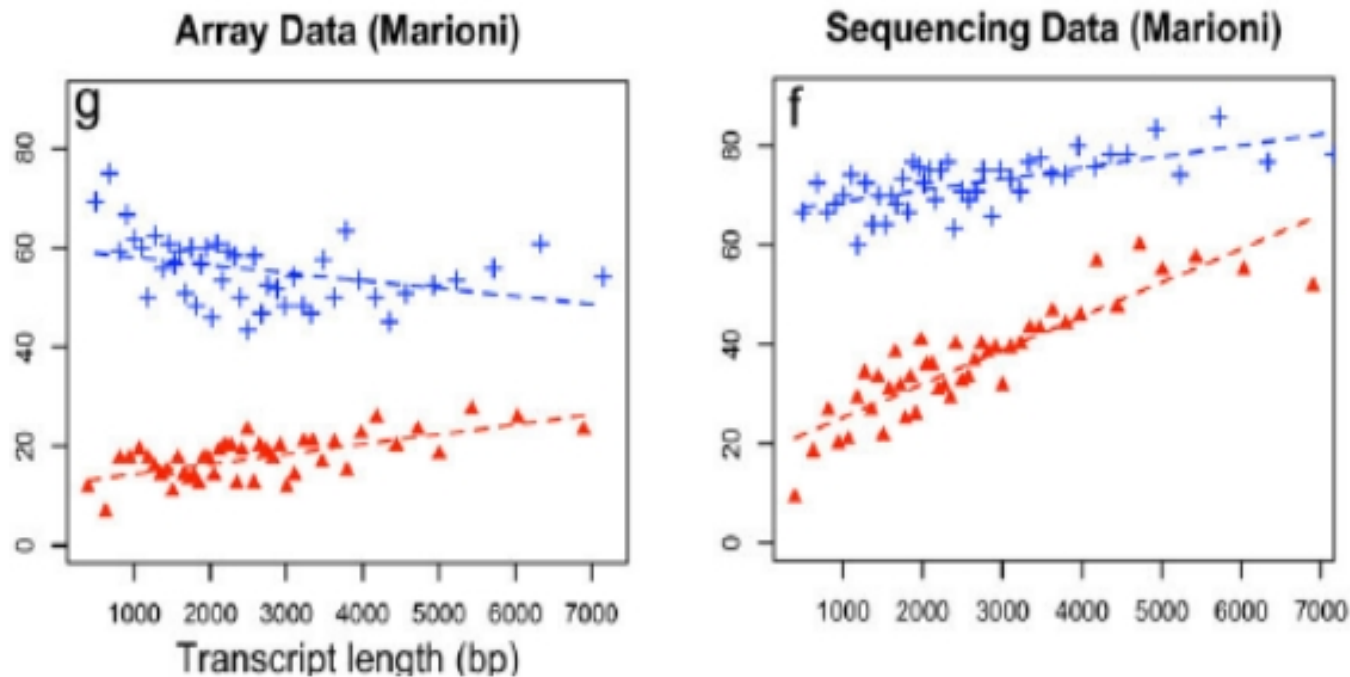
General context

Sequencing Reads



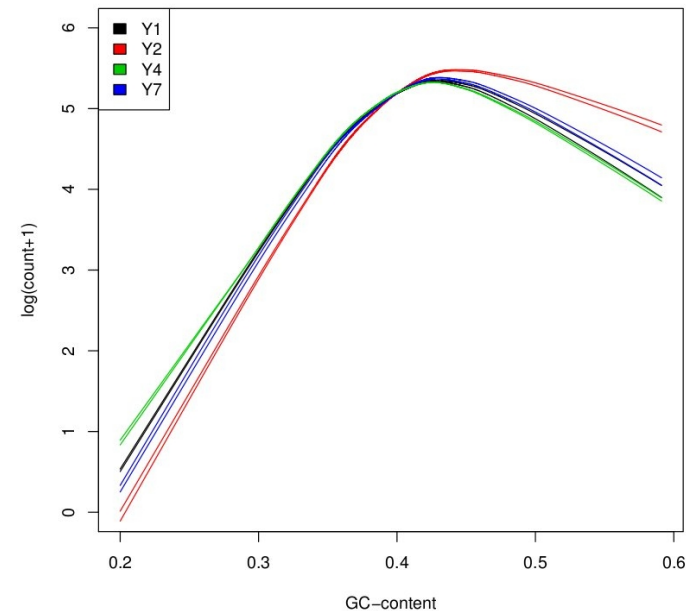
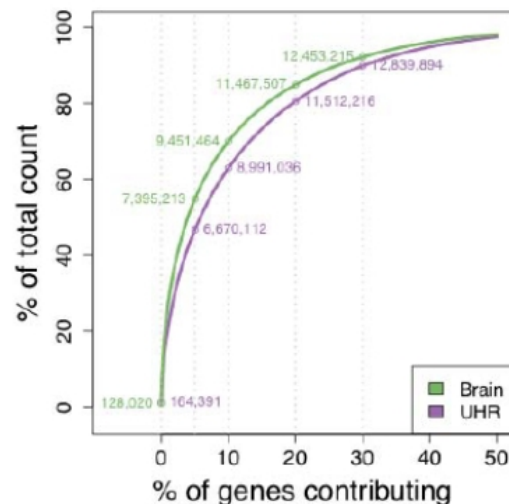
Gene/transcript length dependence

- Counts are proportional to...
 - the transcript length
 - the mRNA expression level.



Count Normalization

- **Transcript length:** *within* library
- **Library size:** *between* libraries
- Many **other biases** ...
 - Differences on the read count distribution among samples.
 - GC content of the gene affects the detection of that gene (Illumina)
 - sequence-specific bias is introduced during the library preparation



Count Normalization

- **RPKM**: Reads Per Kilobase of the transcript per Million mapped reads

$$RPKM = 10^9 \times \frac{C}{N * L}$$

- **C** is the number of mappable reads mapped onto the gene's exons.
- **N** is the total number of mappable reads in the experiment.
- **L** is the total length of the exons in base pairs.
- Fragments Per Kilobase of exon per Million fragments mapped (FPKM),

RNA-Seq Data Analysis Pipeline

Primary

1. Sequence preprocessing



2. Mapping



3. Quantification

Secondary

4. Normalization



5. Differential expression



6. Functional Profiling



Pipeline

RNA-Seq Data Analysis



Babelomics 5

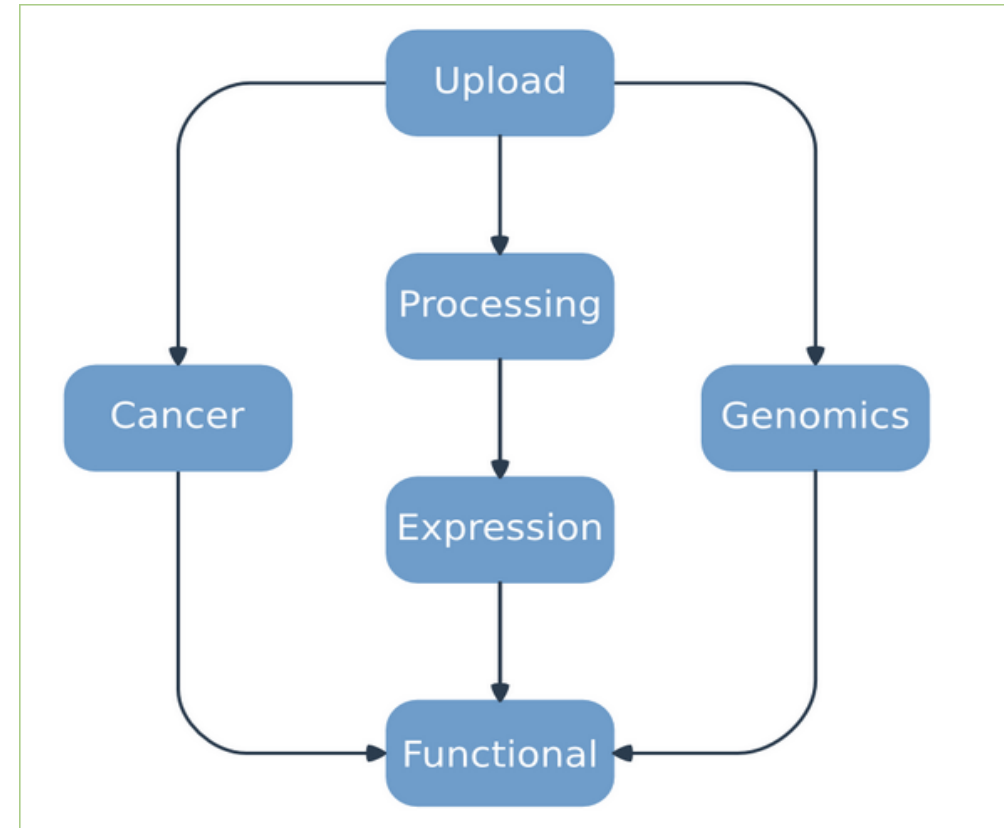
Plataforma de análisis de
datos de Transcriptómica, Proteómica
y Genómica con diferentes abordajes
funcionales

<http://babelomics.bioinfo.cipf.es/>

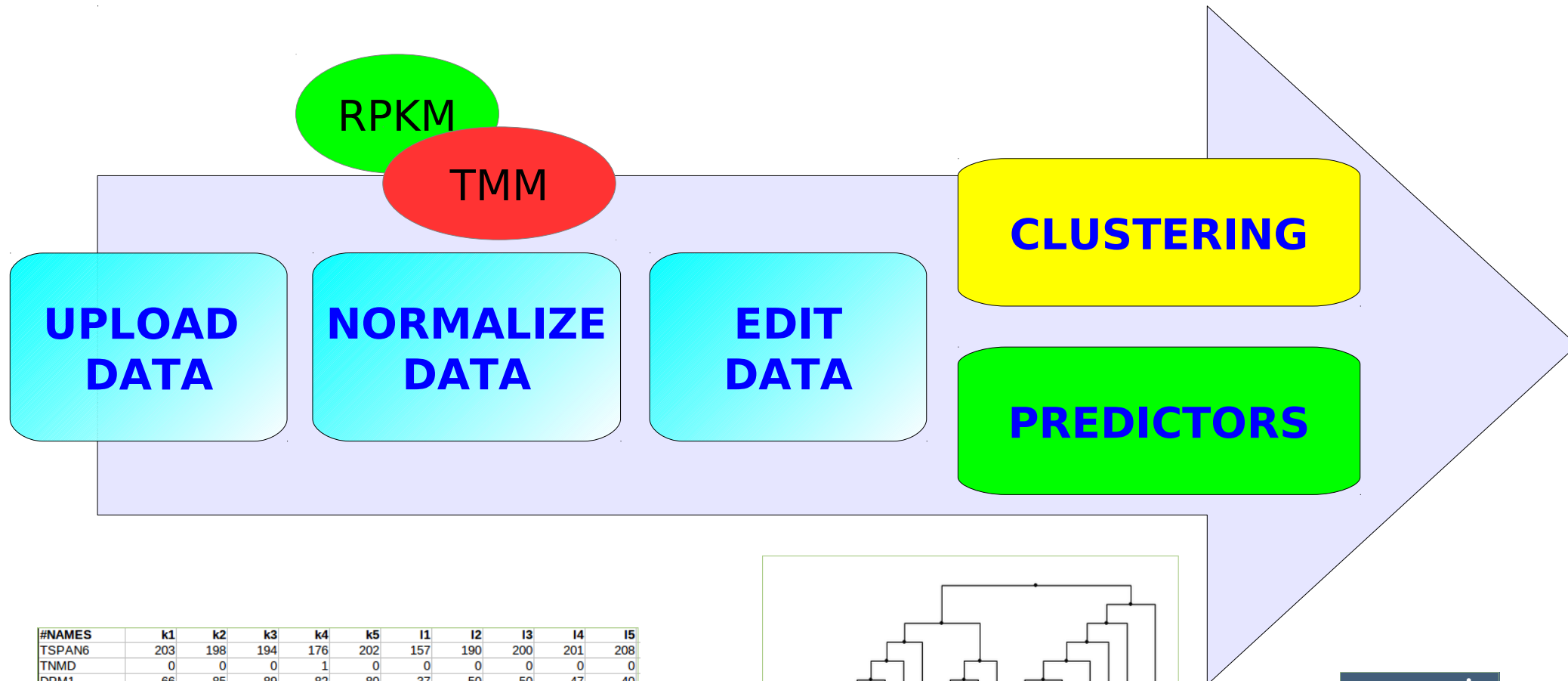
Tool interface

Babelomics 5

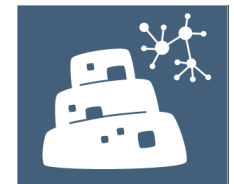
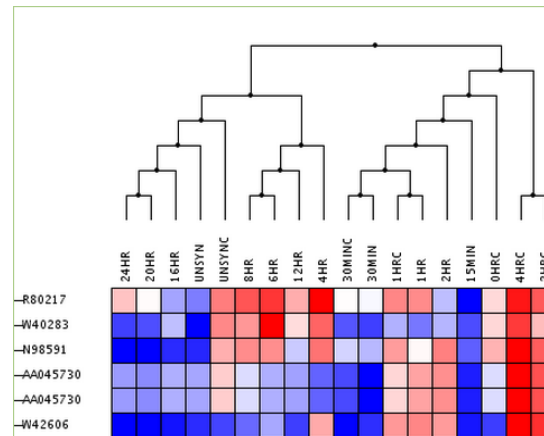
GENE EXPRESSION, GENOME
VARIATION AND FUNCTIONAL
PROFILING ANALYSIS SUITE



Supervised and Unsupervised Classification



#NAMES	k1	k2	k3	k4	k5	l1	l2	l3	l4	l5
TSPAN6	203	198	194	176	202	157	190	200	201	208
TNMD	0	0	0	1	0	0	0	0	0	0
DPM1	66	85	89	82	80	37	50	50	47	40
SCYL3	21	30	31	27	31	28	31	37	15	21
C1orf112	10	12	8	11	18	17	22	12	12	19
FGR	19	28	18	20	10	47	50	43	49	48
FUCA2	240	272	261	256	211	76	82	85	68	83
GCLC	98	100	84	94	86	354	362	373	369	326
NFYA	59	61	53	56	59	59	66	63	66	62
STPG1	34	43	41	31	46	6	7	7	8	7



Differential Expression

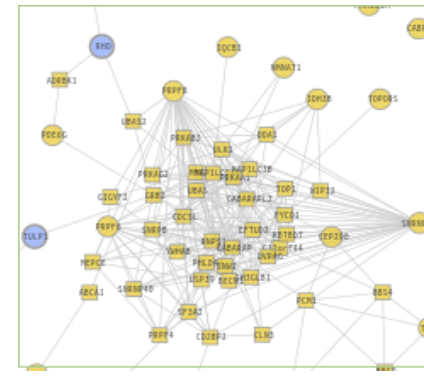
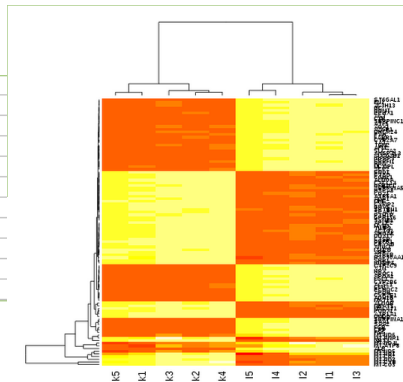
**UPLOAD
DATA**

**EDIT
DATA**

**NORMALIZATION
+
DIFFERENTIAL
EXPRESSION**

**FUNCTIONAL
PROFILING**

#NAMES	k1	k2	k3	k4	k5	l1	l2	l3	l4
TSPAN6	203	198	194	176	202	157	190	200	201
TNMD	0	0	0	1	0	0	0	0	0
DPM1	66	85	89	82	80	37	50	50	47
SCYL3	21	30	31	27	31	28	31	37	15
C1orf112	10	12	8	11	18	17	22	12	12
FGR	19	28	18	20	10	47	50	43	49
FUCA2	240	272	261	256	211	76	82	85	68
GCLC	98	100	84	94	86	354	362	373	369
NFYA	59	61	53	56	59	59	66	63	66
STPG1	34	43	41	31	46	6	7	7	8



Hands on



Babelomics 5

<http://babelomics.bioinfo.cipf.es/>

Processing / Normalization: RNA-Seq
Expression / Differential Expression: RNA-Seq

Online examples

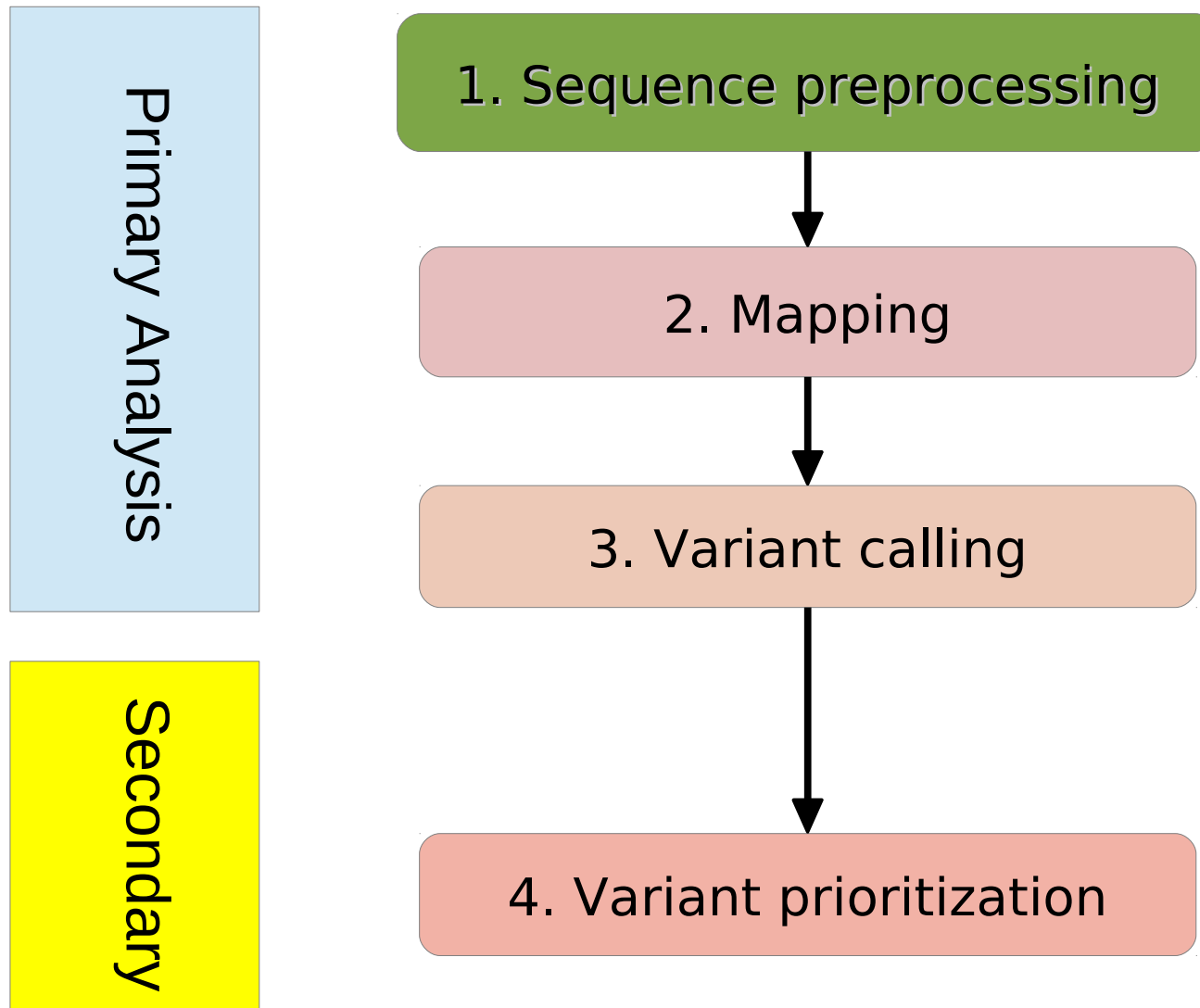
Babelomics 5

RNA-Seq Data Analysis

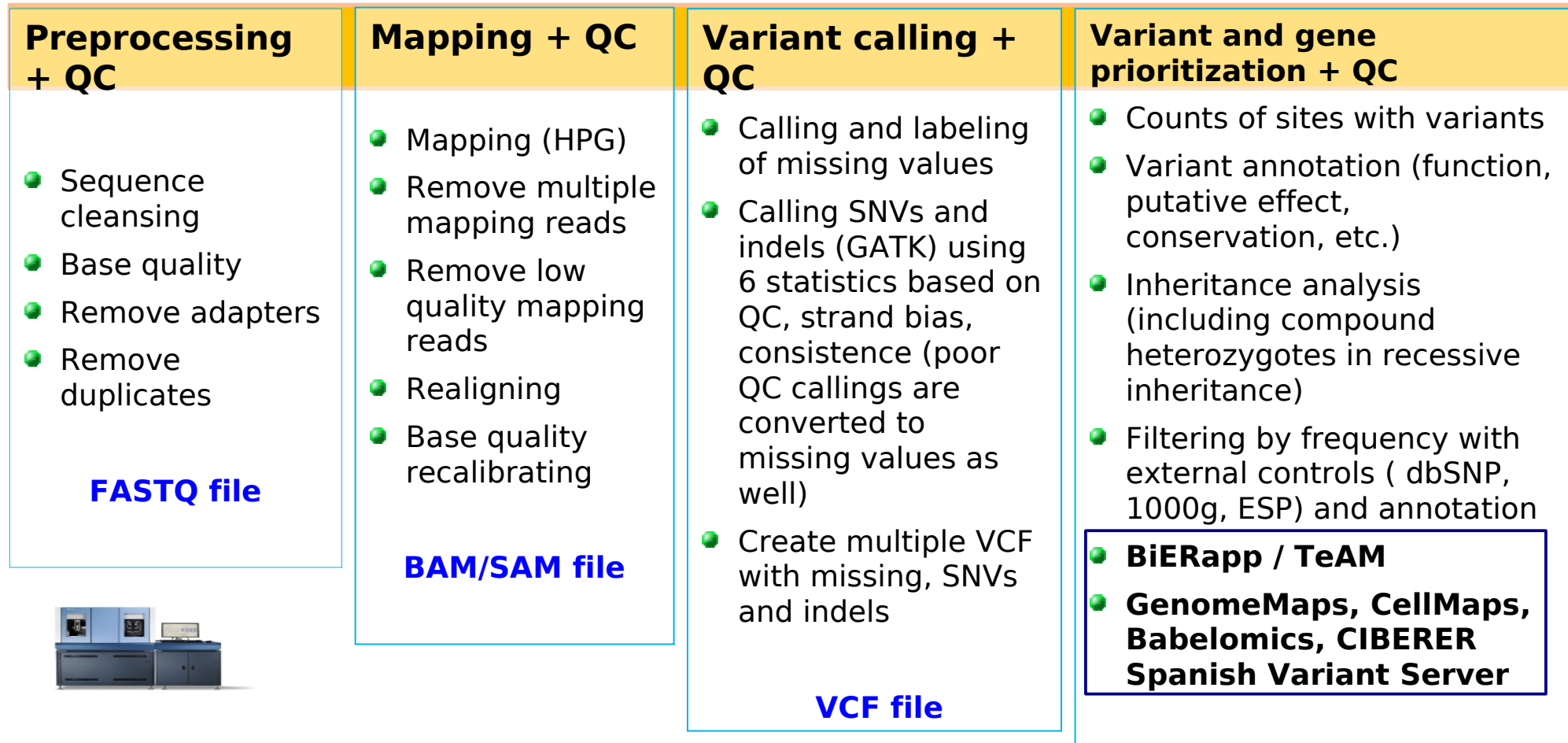
Web Tools for Genomic Data Analysis

- 1) Introduction to NGS Data Analysis
- 2) RNA-Seq Data Analysis
- 3) Resequencing Data Analysis**
 - 1) Pipeline Data Analysis
 - 2) BiERapp (Whole Exome Studies)
 - 3) TEAM (Gene Panel).
 - 4) CSVS (CIBERER Spanish Variant Server), Genome Maps, Cell Maps.
- 4) Omics Data Integration
- 5) Network Analysis

Genomics Data Analysis Pipeline (1)



Genomics Data Analysis Pipeline (2)



Primary analysis

Gene prioritization

Web Tools

Resequencing Data Analysis

How do we prioritize variants in whole exome studies?

<http://bierapp.babelomics.org/>



BiERapp

Discovering variants

Introduction

- Whole-exome sequencing has become a fundamental tool for the discovery of disease-related genes of familial diseases but there are difficulties to **find the causal mutation among the enormous background**
- There are different scenarios, so we need **different and immediate strategies of prioritization**
- Vast amount of **biological knowledge available** in many databases
- We need a tool to **integrate this information and filter immediately** to select candidate variants related to the disease

How does BiERapp work?

Filterings

VCF file
multisample

BiERapp

VARIANT

CellBase

Variant Browser

⌕

Page 1 of 9

🔍

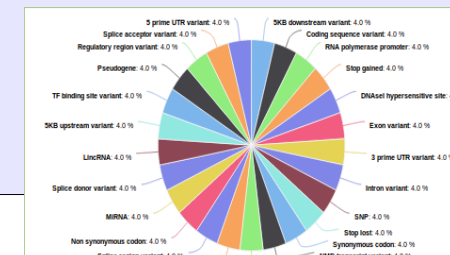
📄

Variant: 1012514058

Variant	Allele	Gene	Samples				S.	Controls (NAF)						EVS	-	-	S.	P.
			NA19000	NA19001	NA19002	NA19003		1000G	1000G-ASR	1000G-ASR	1000G-ASR	1000G-ASR	1000G-ASR					
4102514058	T-C	NPX1	1/1	1/1	1/1	1/1	0.042 (T)	0.002 (T)	0.000 (T)	0.044 (T)	0.089 (T)	0.058	e.	-	-	-	-	
7123047703	T-C	CNDP4	1/1	1/1	1/1	1/1	0.023 (T)	0.025 (T)	0.000 (T)	0.025 (T)	0.000 (T)	0.012	e.	-	-	-	-	
57981270	T-C	HEXB	1/1	1/1	1/1	1/1	0.021 (T)	0.002 (T)	0.000 (T)	0.019 (T)	0.049 (T)	0.031	e.	-	-	-	-	
110795608	T-C	CEL32	1/1	1/1	1/1	1/1	0.070 (T)	0.228 (T)	0.004 (T)	0.038 (T)	0.038 (T)	0.086	e.	-	-	-	-	
1770943990	T-C	SLC39A11	1/1	1/1	1/1	1/1	0.087 (T)	0.344 (T)	0.002 (T)	0.055 (T)	0.001 (T)	0.106	e.	-	-	-	-	
1958879979	C-T	ZNF837	1/1	1/1	1/1	1/1	0.094 (C)	0.132 (C)	0.079 (C)	0.083 (C)	0.073 (C)	0.066	e.	-	-	-	-	
1778289638	A-G	RNF213	1/1	1/1	1/1	1/1	0.000 (A)	0.000 (A)	0.000 (A)	0.000 (A)	0.000 (A)	.	e.	-	-	-	-	
8145795382	T-C	LINC14	1/1	1/1	1/1	1/1	0.008 (T)	0.010 (T)	0.233 (T)	0.089 (T)	0.003 (T)	0.001	S.	-	-	-	-	
1812211090	T-C	DHTRD1	1/1	1/1	1/1	1/1	0.019 (T)	0.077 (T)	0.000 (T)	0.008 (T)	0.000 (T)	0.023	e.	-	-	-	-	
1210572982	A-G	KIRC3	1/1	1/1	1/1	1/1	0.011 (A)	0.047 (A)	0.000 (A)	0.035 (A)	0.000 (A)	0.015	e.	-	-	-	-	

101

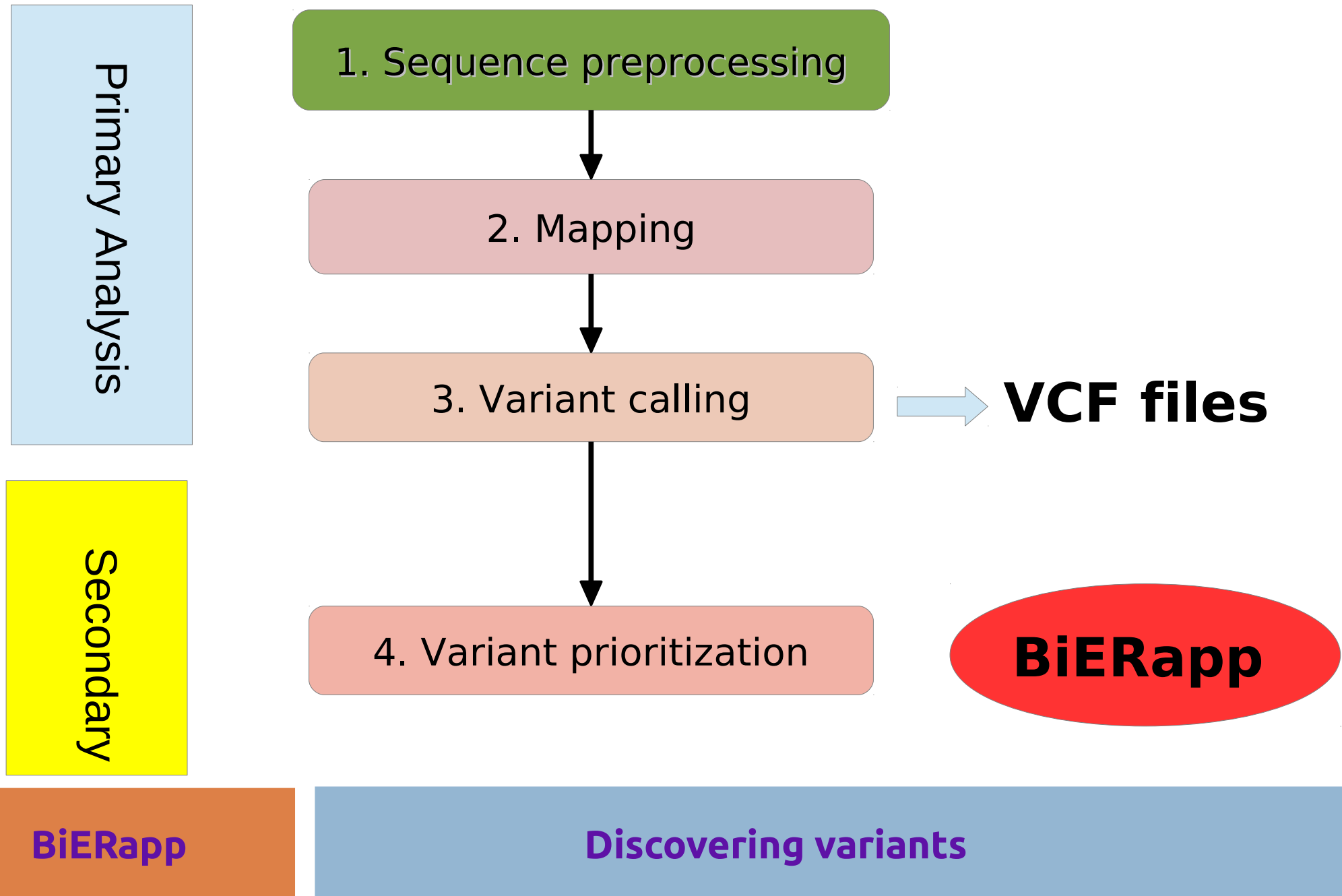
Variant Data



BiERapp

Discovering variants

Input: VCF file



Input: VCF multisample

```
##fileformat=VCFv4.1
##fileDate=20090805
##source=myImputationProgramV3.1
##reference=file:///seq/references/1000GenomesPilot-NCBI36.fasta
##contig=<ID=20,length=62435964,assembly=B36,md5=f126cdf8a6e0c7f379d618ff66beb2da,species="Homo sapiens",taxonomy=x>
##phasing=partial
##INFO=<ID=NS,Number=1,Type=Integer,Description="Number of Samples With Data">
##INFO=<ID=DP,Number=1,Type=Integer,Description="Total Depth">
##INFO=<ID=AF,Number=A,Type=Float,Description="Allele Frequency">
##INFO=<ID=AA,Number=1,Type=String,Description="Ancestral Allele">
##INFO=<ID=DB,Number=0,Type=Flag,Description="dbSNP membership, build 129">
##INFO=<ID=H2,Number=0,Type=Flag,Description="HapMap2 membership">
##FILTER=<ID=q10,Description="Quality below 10">
##FILTER=<ID=s50,Description="Less than 50% of samples have data">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
##FORMAT=<ID=HQ,Number=2,Type=Integer,Description="Haplotype Quality">
#CHROM POS ID REF ALT QUAL FILTER INFO FORMAT NA00001 NA00002 NA00003
20 14370 rs6054257 G A 29 PASS NS=3;DP=14;AF=0.5;DB;H2 GT:GQ:DP:HQ 0|0:48:1:51,51 1|0:48:8:51,51 1/1:43:5:..
20 17330 . T A 3 q10 NS=3;DP=11;AF=0.017 GT:GQ:DP:HQ 0|0:49:3:58,50 0|1:3:5:65,3 0/0:41:3
20 1110696 rs6040355 A G,T 67 PASS NS=2;DP=10;AF=0.333,0.667;AA=T;DB GT:GQ:DP:HQ 1|2:21:6:23,27 2|1:2:0:18,2 2/2:35:4
20 1230237 . T . 47 PASS NS=3;DP=13;AA=T GT:GQ:DP:HQ 0|0:54:7:56,60 0|0:48:4:51,51 0/0:61:2
20 1234567 microsat1 GTC G,GTCT 50 PASS NS=3;DP=9;AA=G GT:GQ:DP 0/1:35:4 0/2:17:2 1/1:40:3
```

**One VCF (Variant Calling Format) file
for family or group**

Getting information

□ SIFT

- SIFT predicts whether an amino acid substitution affects protein function
- **Interpretation:** 1 (tolerated) to 0 (not tolerated)

<http://sift.jcvi.org/>

J. Craig Venter™
INSTITUTE

SIFT

□ PolyPhen

- Polymorphism Phenotyping is a tool which predicts possible impact of an amino acid substitution on the structure and function of a human protein
- **Interpretation:** 1 (probably damage) to 0 (benign)

<http://genetics.bwh.harvard.edu/pph2/index.shtml>



Getting information

e!Ensembl BLAST/BLAT | BioMart | Tools | Downloads | Help & Documentation

Using this website | **Annotation & prediction** | Data access | API & software | About us

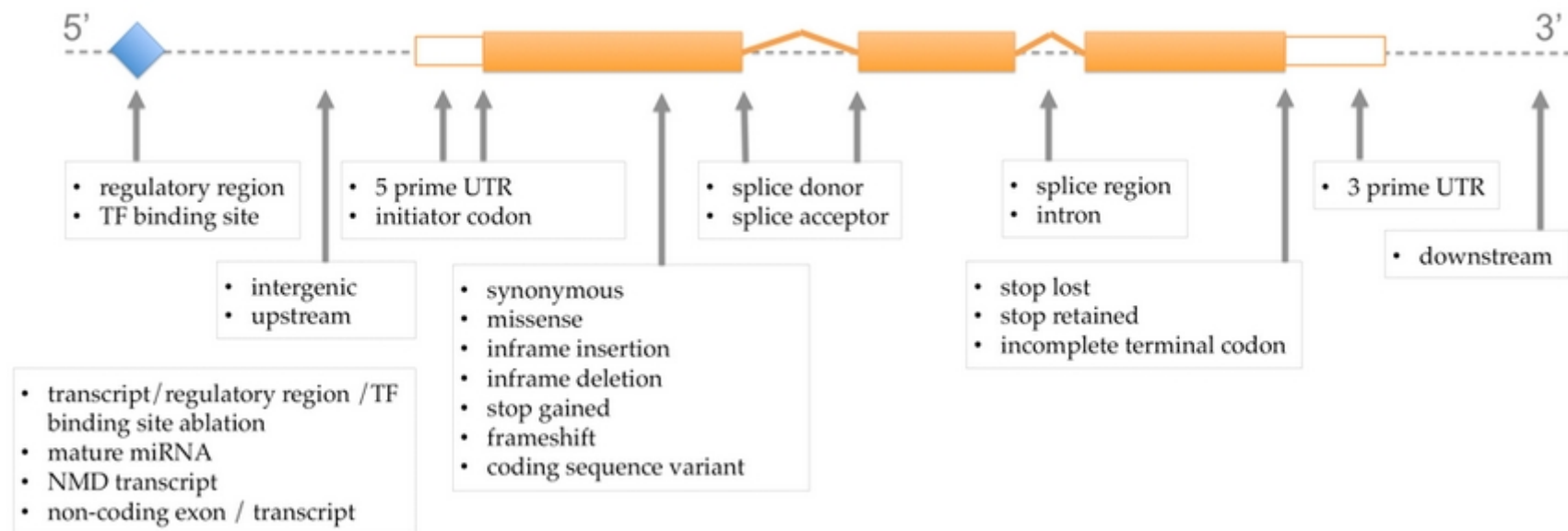
In this section

- Data Description
- Predicted Data
- Import VCF script
- Variation Sources

Home > Help & Documentation > Annotation & Prediction

Ensembl Variation - Predicted data


Consequence type or effect



http://www.ensembl.org/info/genome/variation/predicted_data.html

Tool interface

<http://bierapp.babelomics.org/>

[Menu](#) BierApp  [Home](#)

Overview

Welcome to the gene/variant prioritization tool of the BIER (the Team of Bioinformatics for Rare Diseases). This interactive tool allows finding genes affected by deleterious variants that segregate along family pedigrees, case-controls or sporadic samples.







Try an Example

Here you can try all the filtering options and discover the gene affected in a test family.

Analyze your own families or case-control data

Here you can upload your VCF file containing the exomes to be analyzed. Define the thresholds of allele frequencies, pathogenicity, conservation; the type of variants sought; and define the type of inheritance and the segregation schema along the family.

Supported by



[logout](#) [upload & manage](#) [profile](#) [jobs](#) [support](#)

BiERapp

Discovering variants

Tool interface

Menu BierApp Home

Example 1000G (Short)

Filter

Clear Submit

Segregation

	0/0	0/1	1/1
NA19600:	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
NA19660:	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
NA19661:	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
NA19685:	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

MAF

1000G MAF <: 0.1

EVS MAF <:

1000G Populations

African MAF <:

American MAF <:

Asian MAF <:

European MAF <:

Position

Consequence Type

- ☐ SKB_downstream_variant
- ☐ coding_sequence_variant
- ☐ RNA_polymerase_promoter
- ☐ stop_gained
- ☐ DNaseI_hypersensitive_site
- ☒ exon_variant
- ☐ 3_prime_UTR_variant
- ☐ intron_variant
- ☐ SNP
- ☐ stop_lost
- ☐ synonymous_codon
- ☐ NMD_transcript_variant
- ☐ CpG_island
- ☐ miRNA_target_site

Variant Browser

Page 1 of 9

Variant	Alleles	Gene	Samples				...	Controls (MAF)						Ph
			NA19600	NA19660	NA19661	NA19685		1000G	1000G-AFR	1000G-ASI	1000G-AME	1000G-EUR	EVS				
4:103514658	T>C	NFKB1	1/1	1/1	1/1	1/1		0.042(T)	0.002(T)	0.000(T)	0.064(T)	0.089(T)	0.058	e...	.	.	
7:135047703	T>C	CNOT4	1/1	1/1	1/1	1/1		0.013(T)	0.055(T)	0.000(T)	0.005(T)	0.000(T)	0.012	e...	.	.	
5:73981270	T>C	HEXB	1/1	1/1	1/1	1/1		0.021(T)	0.002(T)	0.000(T)	0.019(T)	0.049(T)	0.031	e...	0...	0...	
1:109795608	T>C	CELSR2	1/1	1/1	1/1	1/1		0.070(T)	0.228(T)	0.004(T)	0.036(T)	0.036(T)	0.086	e...	1...	.	
17:70943990	T>C	SLC39A11	1/1	1/1	1/1	1/1		0.087(T)	0.344(T)	0.002(T)	0.055(T)	0.001(T)	0.106	e...	0...	0...	
19:58879976	C>T	ZNF837	1/1	1/1	1/1	1/1		0.094(C)	0.152(C)	0.079(C)	0.083(C)	0.073(C)	0.066	e...	0...	0...	
17:78298938	A>G	RNF213	1/1	1/1	1/1	1/1		0.000(A)	0.000(A)	0.000(A)	0.000(A)	0.000(A)	.	e...	0...	1...	
8:145745182	T>C	LRRC14	1/1	1/1	1/1	1/1		0.068(T)	0.010(T)	0.203(T)	0.069(T)	0.003(T)	0.001	5...	0...	.	
10:12111090	T>C	DHTRD1	1/1	1/0	1/1	0/1		0.019(T)	0.077(T)	0.000(T)	0.008(T)	0.000(T)	0.033	e...	0...	0...	

Variant Data

Genomic Context Effect & Annotation Study Summary

Effects

Consequence type

Num variants: 1000 Samples

Num samples: 4 NA19600

Num indels: 21 NA19660

Num biallelic: 1000 NA19661

Num multiallelic: 0 NA19685

Num transitions: 748

Num transversions: 231

% PASS: 100%

Ti/Tv Ratio: 3.24

Avg Quality: 106.90

3 prime UTR variant: 4.0 %

SKB downstream variant: 4.0 %

Coding sequence variant: 4.0 %

RNA polymerase promoter: 4.0 %

Stop gained: 4.0 %

DNaseI hypersensitive site: 4.0 %

Exon variant: 4.0 %

3 prime UTR variant: 4.0 %

Intron variant: 4.0 %

SNP: 4.0 %

Stop lost: 4.0 %

Synonymous codon: 4.0 %

NMD transcript variant: 4.0 %

CpG island: 4.0 %

MIRNA target site: 4.0 %

Splice region variant: 4.0 %

Non synonymous codon: 4.0 %

MIRNA: 4.0 %

LincRNA: 4.0 %

SKB upstream variant: 4.0 %

TF binding site variant: 4.0 %

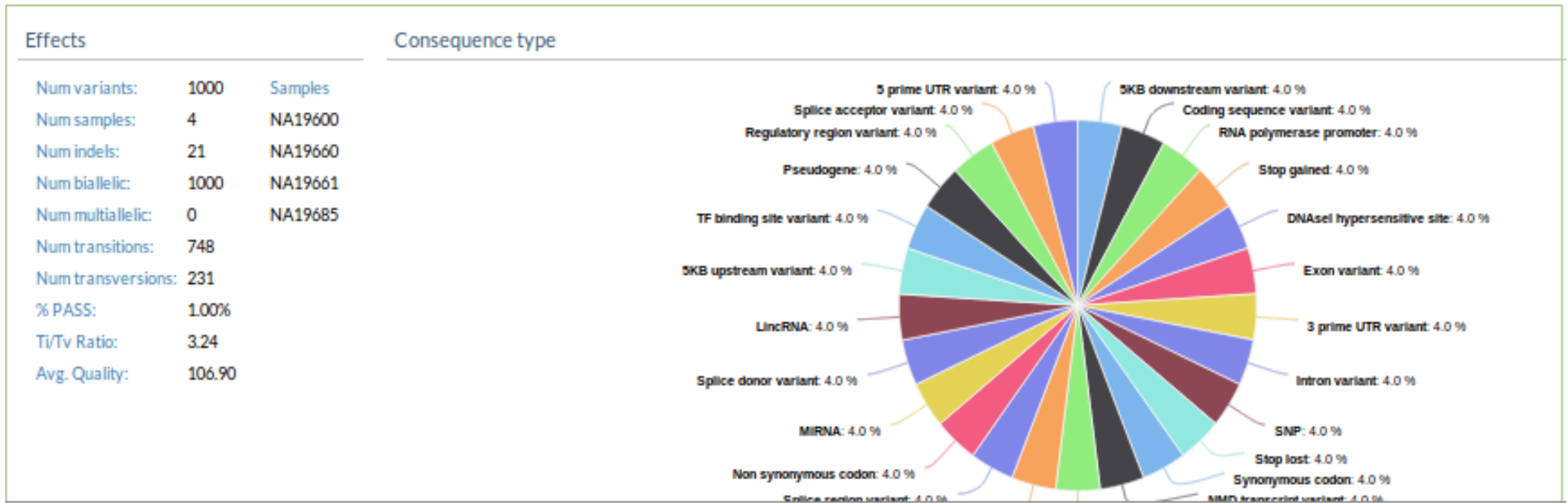
Pseudogene: 4.0 %

Regulatory region variant: 4.0 %

Splice acceptor variant: 4.0 %

Results

1. Summary. Description about number of variants, INDELs... Also a distribution of consequences types.



Results

2. List of candidate variants.

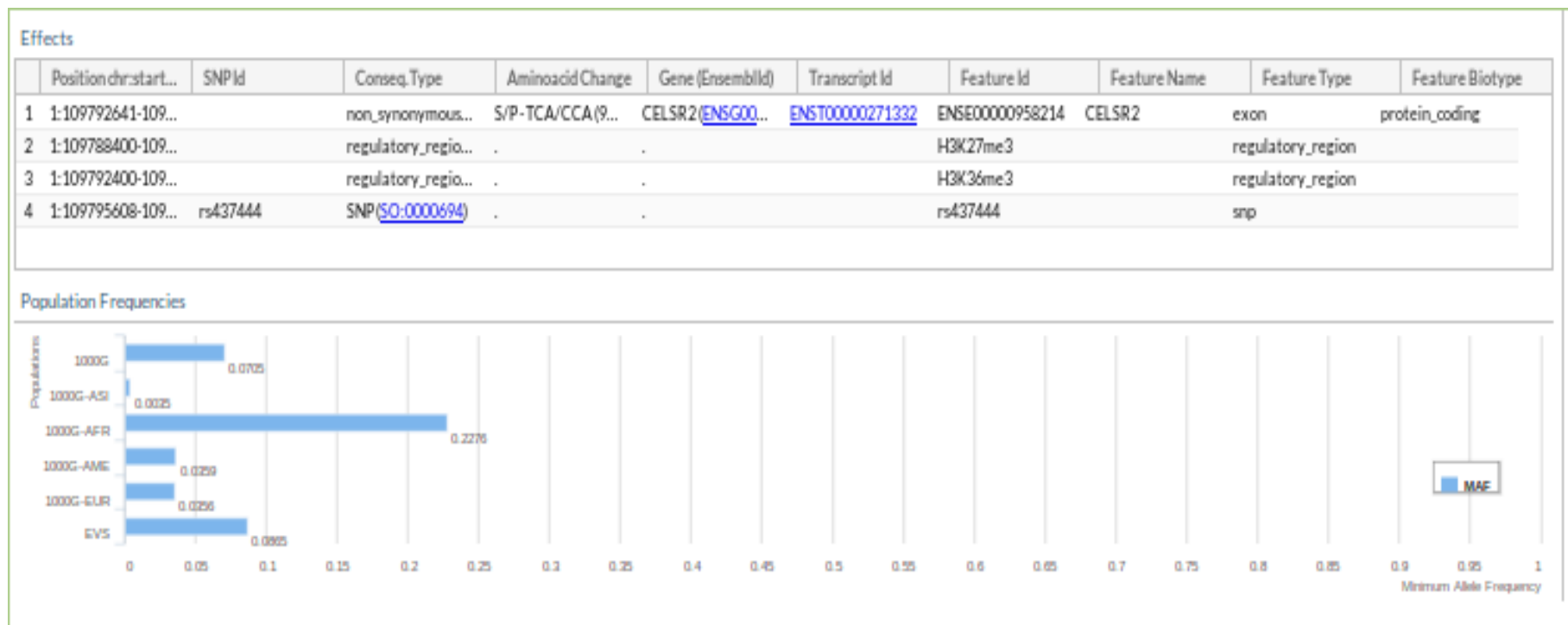
We can order this list by several criteria.

Variant Browser																	
« Page 1 of 9 » C																	
Variant	Alleles	Gene	Samples				S..	Controls (MAF)						S..	Ph
			NA19600	NA19660	NA19661	NA19685		1000G	1000G-AFR	1000G-ASI	1000G-AME	1000G-EUR	EVS				
4:103514658	T>C	NFKB1	1/1	1/1	1/1	1/1		0.042 (T)	0.002 (T)	0.000 (T)	0.064 (T)	0.089 (T)	0.058	e..	.	.	
7:135047703	T>C	CNOT4	1/1	1/1	1/1	1/1		0.013 (T)	0.055 (T)	0.000 (T)	0.005 (T)	0.000 (T)	0.012	e..	.	.	
5:73981270	T>C	HEXB	1/1	1/1	1/1	1/1		0.021 (T)	0.002 (T)	0.000 (T)	0.019 (T)	0.049 (T)	0.031	e..	Q..	Q..	
1:109795608	T>C	CELSR2	1/1	1/1	1/1	1/1		0.070 (T)	0.228 (T)	0.004 (T)	0.036 (T)	0.036 (T)	0.086	e..	1..	.	
17:70943990	T>C	SLC39A11	1/1	1/1	1/1	1/1		0.087 (T)	0.344 (T)	0.002 (T)	0.055 (T)	0.001 (T)	0.106	e..	Q..	Q..	
19:58879976	C>T	ZNF837	1/1	1/1	1/1	1/1		0.094 (C)	0.152 (C)	0.079 (C)	0.083 (C)	0.073 (C)	0.066	e..	Q..	Q..	
17:78298938	A>G	RNF213	1/1	1/1	1/1	1/1		0.000 (A)	0.000 (A)	0.000 (A)	0.000 (A)	0.000 (A)	.	e..	Q..	1..	
8:145745182	T>C	LRRC14	1/1	1/1	1/1	1/1		0.068 (T)	0.010 (T)	0.203 (T)	0.069 (T)	0.003 (T)	0.001	5..	Q..	.	
10:12111090	T>C	DHTKD1	1/1	1/0	1/1	0/1		0.019 (T)	0.077 (T)	0.000 (T)	0.008 (T)	0.000 (T)	0.033	e..	Q..	Q..	
12:10572982	A>G	KLRC3	1/1	1/1	1/1	1/1		0.011 (A)	0.043 (A)	0.000 (A)	0.005 (A)	0.000 (A)	0.015	e..	.	.	
« »																	
Variant Data																	

Results

3. Effects for each transcript where we detected a candidate variant.

The plot shows MAFs for different groups (1000 Genomes, Exome Variant Server)



Results

4. Visualization of candidate variants from GenomeMaps

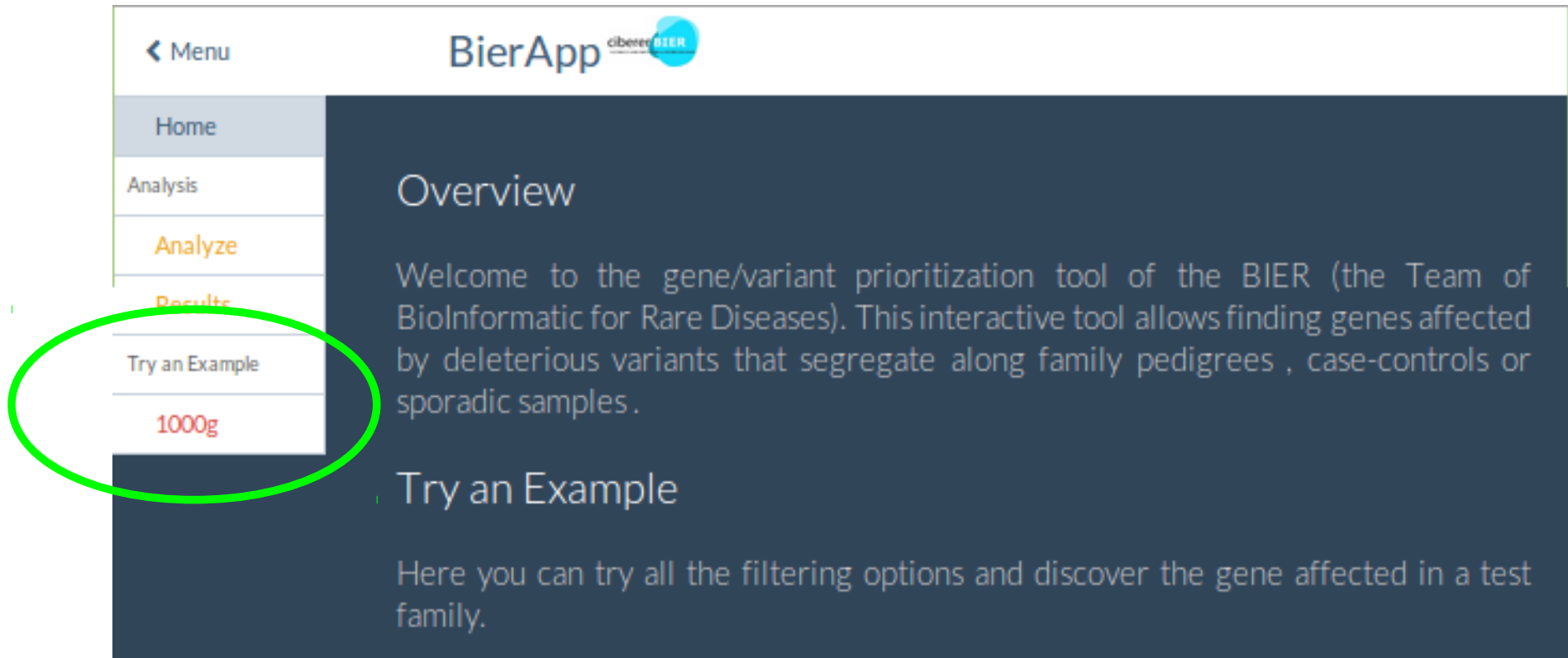


Remarks

- The proposed web-based interactive framework has **great potential to detect disease-related variants** in familial diseases as demonstrated by its successful use in several studies
- **The use of the filters is interactive** and the results are almost instantaneously displayed in a panel that includes the genes affected, the variants and specific information for them
- Candidate variants are **new knowledge useful for future diagnostic**

Hands on

<http://bierapp.babelomics.org/>

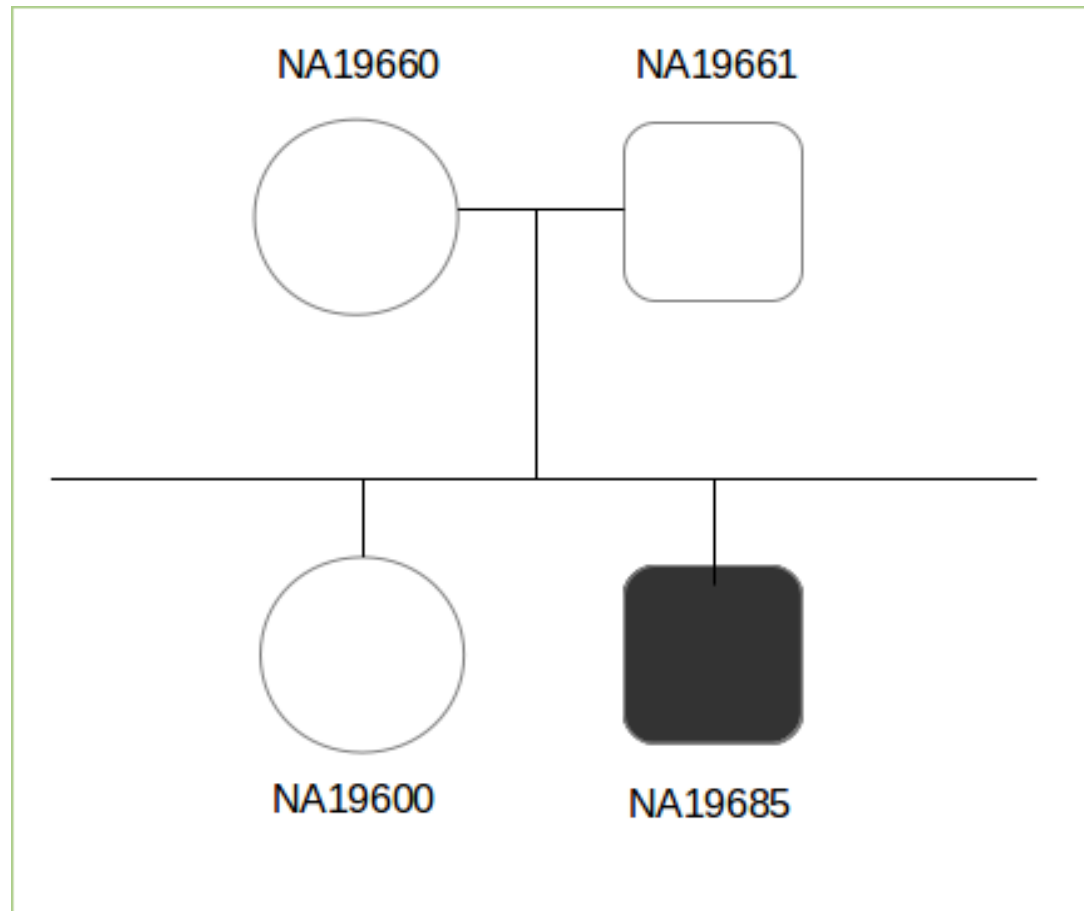


BiERapp

Discovering variants

Hands on

Pedigree



Hands on

Case 1.

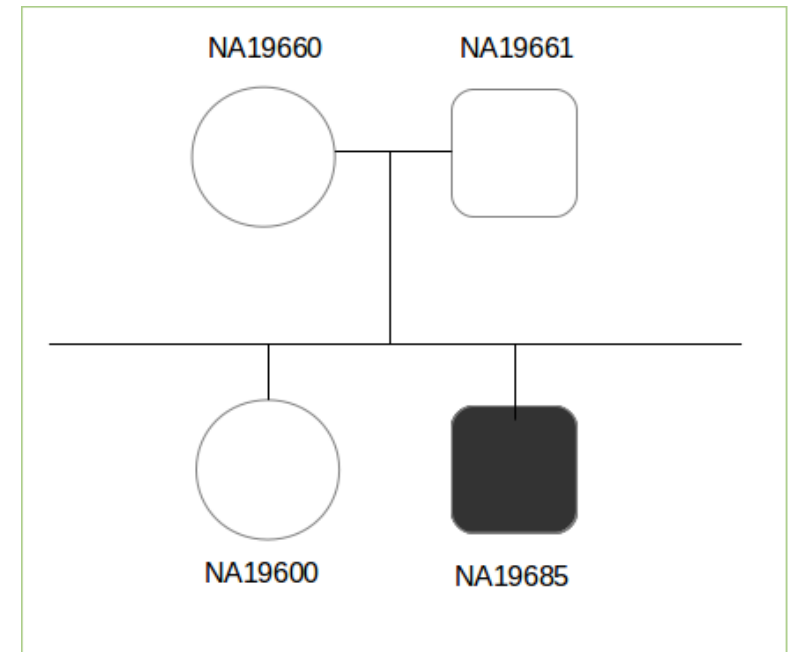
- De novo variants

How many variants?

Case 2.

- Recessive heritage

How many variants?

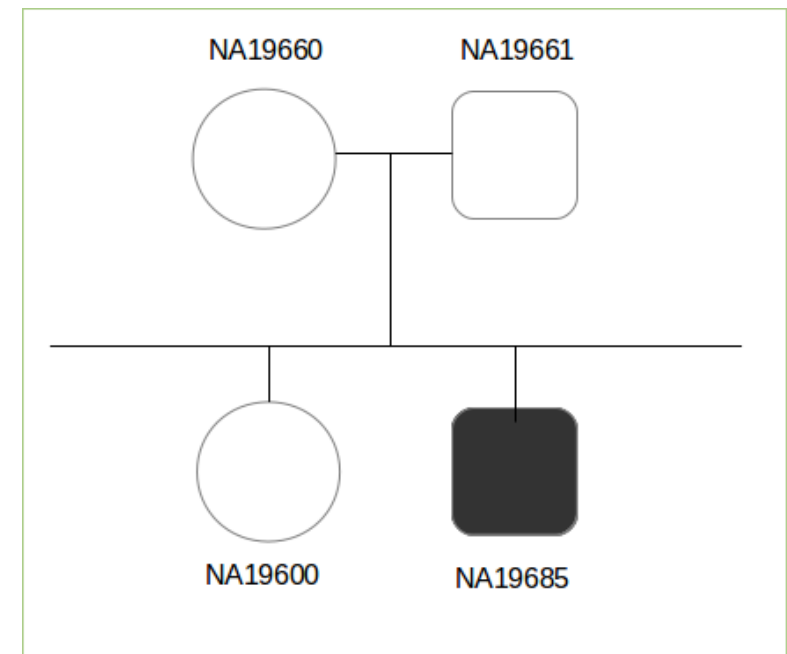


Hands on

Case 3.

- Recessive heritage
- Rare disease (MAF < 0.1)

How many variants?



Case 4.

- Variants in mother and daughter at the same time

How many variants?

Hands on

Case 5.

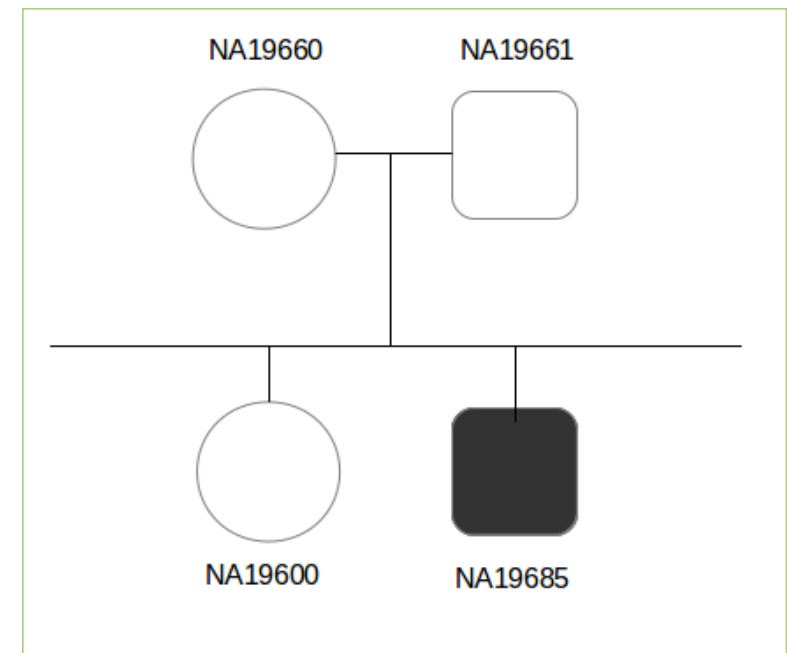
- Variants in mother and daughter at the same time
- Only in chromosome 4

How many variants?

Case 6.

- Variants in mother and daughter at the same time
- Only in these genes: HEXB, NFKB1, KLRC3

How many variants?



More information

Nucleic Acids Research Advance Access published May 6, 2014

Nucleic Acids Research, 2014 **1**
doi: 10.1093/nar/gku407

A web-based interactive framework to assist in the prioritization of disease candidate genes in whole-exome sequencing studies

Alejandro Alemán^{1,2}, Francisco Garcia-Garcia¹, Francisco Salavert^{1,2}, Ignacio Medina¹ and Joaquín Dopazo^{1,2,3,*}

¹Computational Genomics Department, Centro de Investigación Príncipe Felipe (CIPF), Valencia 46012, Spain,

²Bioinformatics of Rare Diseases (BIER), CIBER de Enfermedades Raras (CIBERER), Valencia 46010, Spain and

³Functional Genomics Node, (INB) at CIPF, Valencia 46012, Spain



BiERapp Tutorial:

<http://bierapp.babelomics.org/>



BiERapp

Discovering variants

Can I interpret sequencing data for diagnostic?

<http://team.babelomics.org/>



TEAM

Targeted Enrichment Analysis and Management

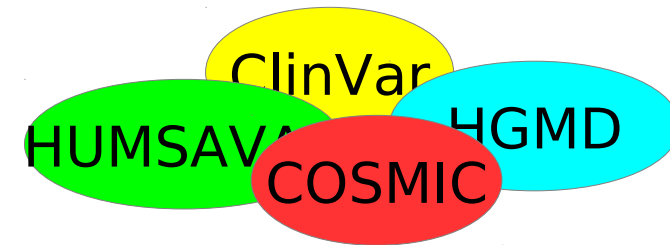
Introduction

Sequencing
data

Biological
knowledge



TEAM

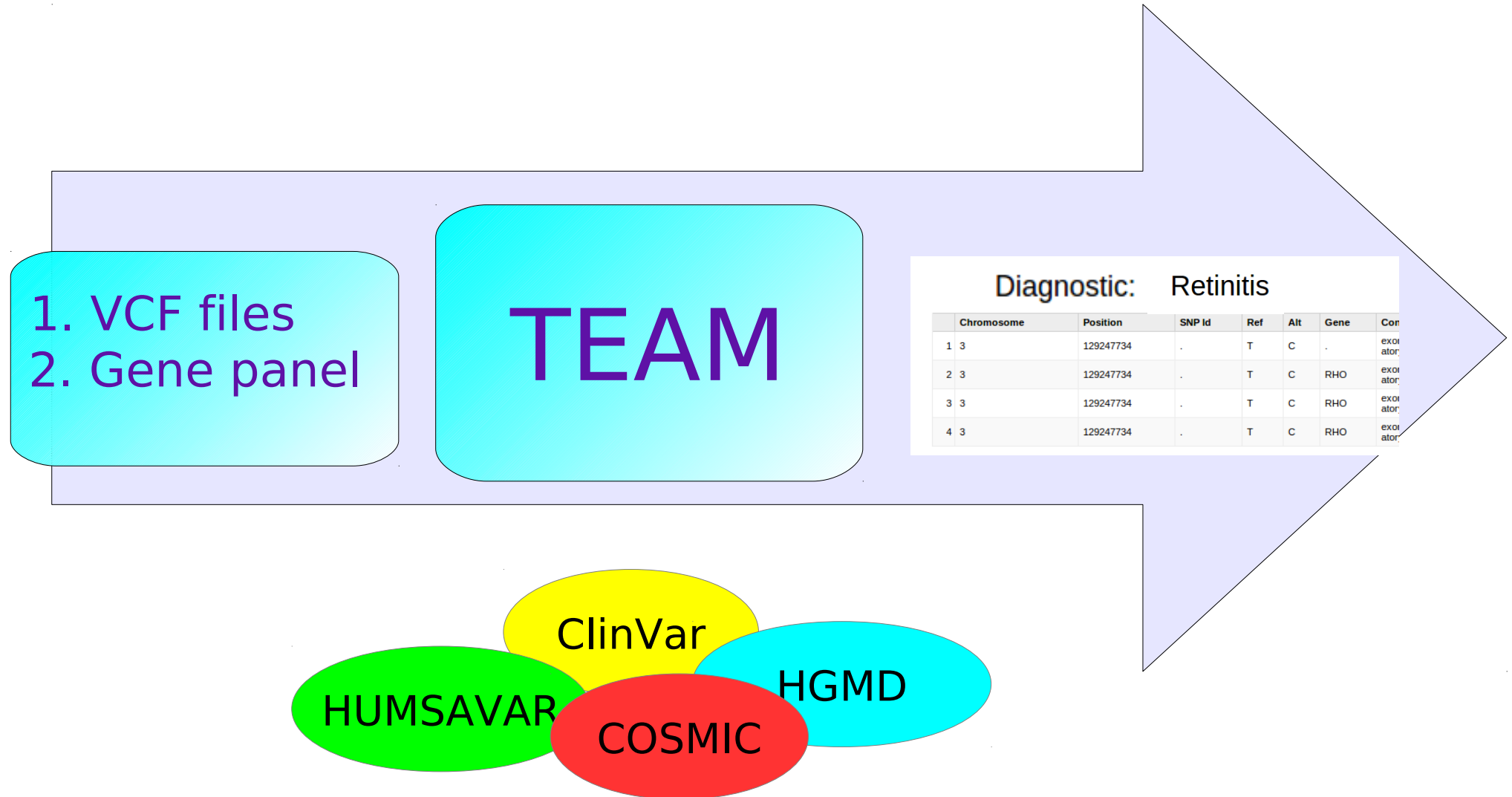


Diagnostic

TEAM

Targeted Enrichment Analysis and Management

How does TEAM work?



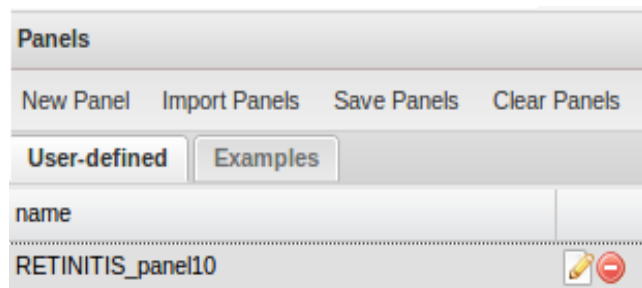
TEAM

Targeted Enrichment Analysis and Management

How does TEAM work?

<http://team.babelomics.org/>

1. Defining panel



Panels

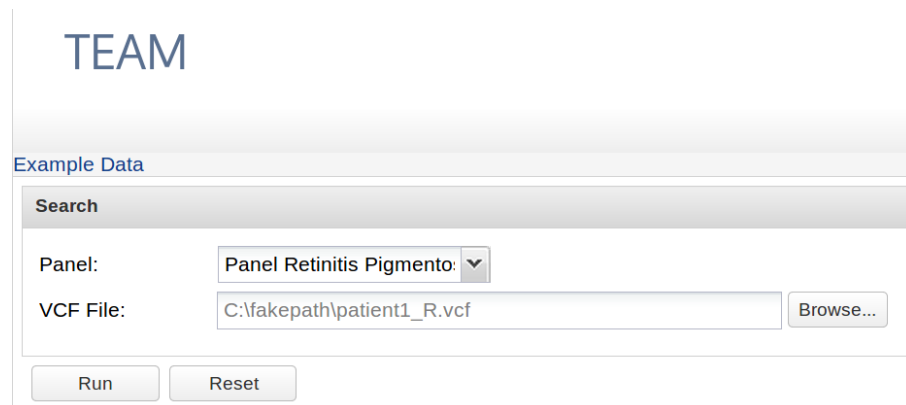
New Panel Import Panels Save Panels Clear Panels

User-defined Examples

name

RETINITIS_panel10

2. Uploading input data



TEAM

Example Data

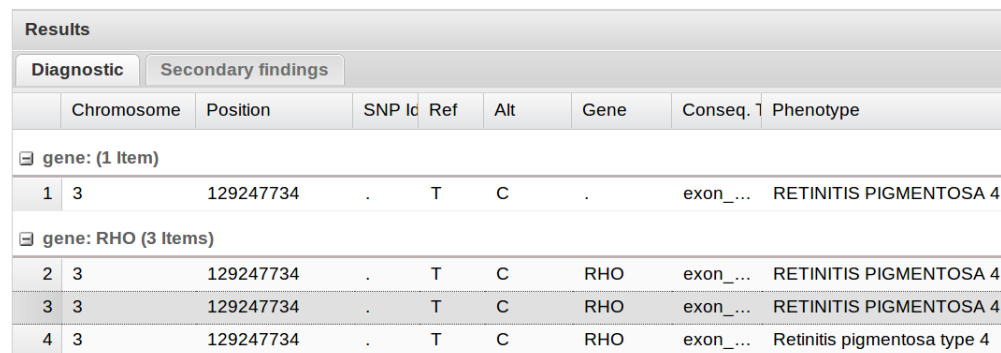
Search

Panel: Panel Retinitis Pigmento: ▾

VCF File: C:\fakepath\patient1_R.vcf Browse...

Run Reset

3. Getting results



Results							
Diagnostic Secondary findings							
	Chromosome	Position	SNP Id	Ref	Alt	Gene	Conseq. 1 Phenotype
gene: (1 item)							
1	3	129247734	.	T	C	.	exon_... RETINITIS PIGMENTOSA 4
gene: RHO (3 items)							
2	3	129247734	.	T	C	RHO	exon_... RETINITIS PIGMENTOSA 4
3	3	129247734	.	T	C	RHO	exon_... RETINITIS PIGMENTOSA 4
4	3	129247734	.	T	C	RHO	exon_... Retinitis pigmentosa type 4

TEAM

Targeted Enrichment Analysis and Management

How to define a panel?

1. Name of panel

2. Diseases

3. Adding:
- more genes
- mutations

4. Save panel

The screenshot shows the 'Panel Manager' window with the following components and annotations:

- Name:** A text field containing 'RETINITIS_10'. An arrow points to it from the text '1. Name of panel'.
- Diseases (Drag):** A list of diseases. An arrow points to 'RETINITIS PIGMENTOSA 14' from the text '2. Diseases'.
- Primary Disease (Drop):** A list containing 'RETINITIS PIGMENTOSA 10', 'RETINITIS PIGMENTOSA 13', and 'RETINITIS PIGMENTOSA 20'.
- Genes:** A list containing 'IMPDH1', 'PRPF8', and 'RPE65'. Each gene has a red minus button to its right. An arrow points to this section from the text '3. Adding: - more genes - mutations'.
- Mutations:** A table with columns 'Chr', 'Pos', 'Ref', 'Alt', and 'Gene'. An arrow points to the 'Add Mutation' button below it from the text '3. Adding: - more genes - mutations'.
- Text/Bed File:** Radio buttons for 'Text' (selected) and 'Bed File'. Below is a text field containing 'BRCA2,PPL'. An arrow points to this section from the text '3. Adding: - more genes - mutations'.
- Add Genes:** A button. An arrow points to it from the text '3. Adding: - more genes - mutations'.
- Save panel:** An arrow points from the text '4. Save panel' to the 'Add new panel' button at the bottom right.
- Other controls:** At the bottom left, there are fields for 'PolyPhen:' and 'Sift:' with up/down arrows. At the bottom right, there are buttons for 'Add new panel', 'Clear', and 'Close'.

How to define a panel?

Add mutation

Chr: 8 Pos: 55539395 Ref: A Alt: T Gene Name: RP1 Disease Name: Lung cancer 2

Reset Check Add Mutation

Region overview Window size: 583 nts

55,539,104 55,539,395 55,539,686

Sequence: AAGCACATAACTAAAATTGCCGTTTGACAGGAGATAATCTATGTAAAGAGGGAGATAAGTCTTTT

Gene

SNP: P_ESP_8_55539357, 8_55539353, rs202016292, rs58051614, rs201613551, rs202226256, rs200135800, rs2293869, COSM486527, rs202057087

T 8:55,539,394 Genome Viewer

Adding
new mutations

Checking
mutations from
Genome Viewer

TEAM

Targeted Enrichment Analysis and Management

Results

Results								
Diagnostic Secondary findings								
	Chromosome	Position	SNP Id	Ref	Alt	Gene	Conseq. Type	Phenotype
gene: (1 item)								
1	3	129247734	.	T	C	.	exon_vari...	RETINITIS PIGMENTOSA 4
gene: RHO (3 items)								
2	Variant Effect - 3:129247734 T>C							
3		Position chr:start:end (strand)	SNP Id			Conseq. Type		Aminoacid Change
4	1	3:129247734-129247734 (+)	CM920608			SNP (SO:0000694)		.
	2	3:129247483-129247937 (+)				synonymous_codon (SO:00...		P/P - CCC/CCC (53)
	3	3:129245550-129248350				regulatory_region_variant (...)		.
	4	3:129247734-129247734 (+)	rs28933395			SNP (SO:0000694)		.

A. Web results

B. PDF report

Diagnostic: Retinitis

	Chromosome	Position	SNP Id	Ref	Alt	Gene	Con
1	3	129247734	.	T	C	.	exon ator
2	3	129247734	.	T	C	RHO	exon ator
3	3	129247734	.	T	C	RHO	exon ator
4	3	129247734	.	T	C	RHO	exon ator

TEAM

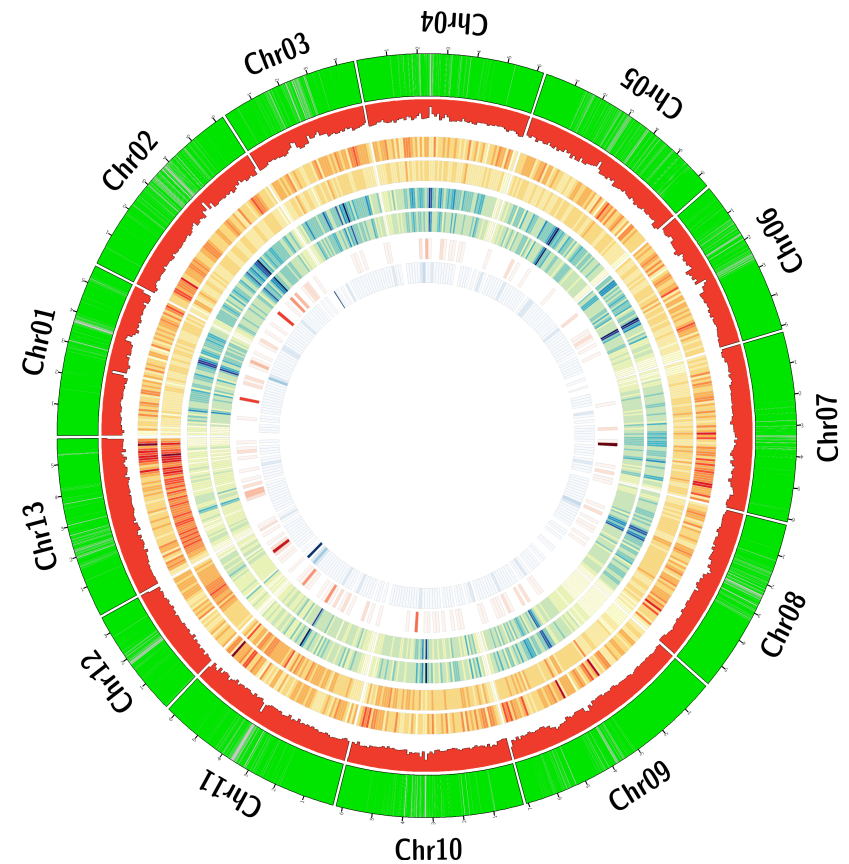
Targeted Enrichment Analysis and Management

Remarks

- TEAM is a **free** web tool
- **Easy**-to-use and **powerful**
- TEAM helps you for **diagnostic**

Next improvements:

- Inclusion of a **database with public panels genes** of various diseases.
- **Comparative Analysis** for groups of panels.
- **Visualization results.**



Hands on

<http://team.babelomics.org/>

- 1) Download **example data** from TEAM (3 VCF files).
- 2) **Select the panel** for Retinitis Pigmentosa and **evaluate all three samples**. Do you have variants related to Retinitis for each of the three patients?
- 3) **Generate a PDF report** for each patient including variants related to diagnostic and secondary findings.
- 4) **Design a new panel** for Usher disease.

More information

Nucleic Acids Research Advance Access published May 26, 2014

Nucleic Acids Research, 2014 **1**

doi: 10.1093/nar/gku472

A web tool for the design and management of panels of genes for targeted enrichment and massive sequencing for clinical applications

Alejandro Alemán^{1,2}, Francisco Garcia-Garcia¹, Ignacio Medina¹ and Joaquín Dopazo^{1,2,3,*}

¹Computational Genomics Department, Centro de Investigación Príncipe Felipe (CIPF), Valencia, 46012, Spain,

²Bioinformatics of Rare Diseases (BIER), CIBER de Enfermedades Raras (CIBERER), Valencia, 46012, Spain and

³Functional Genomics Node, (INB) at CIPF, Valencia, 46012, Spain



TEAM Tutorial:

<http://ciberer.es/bier/team>



TEAM

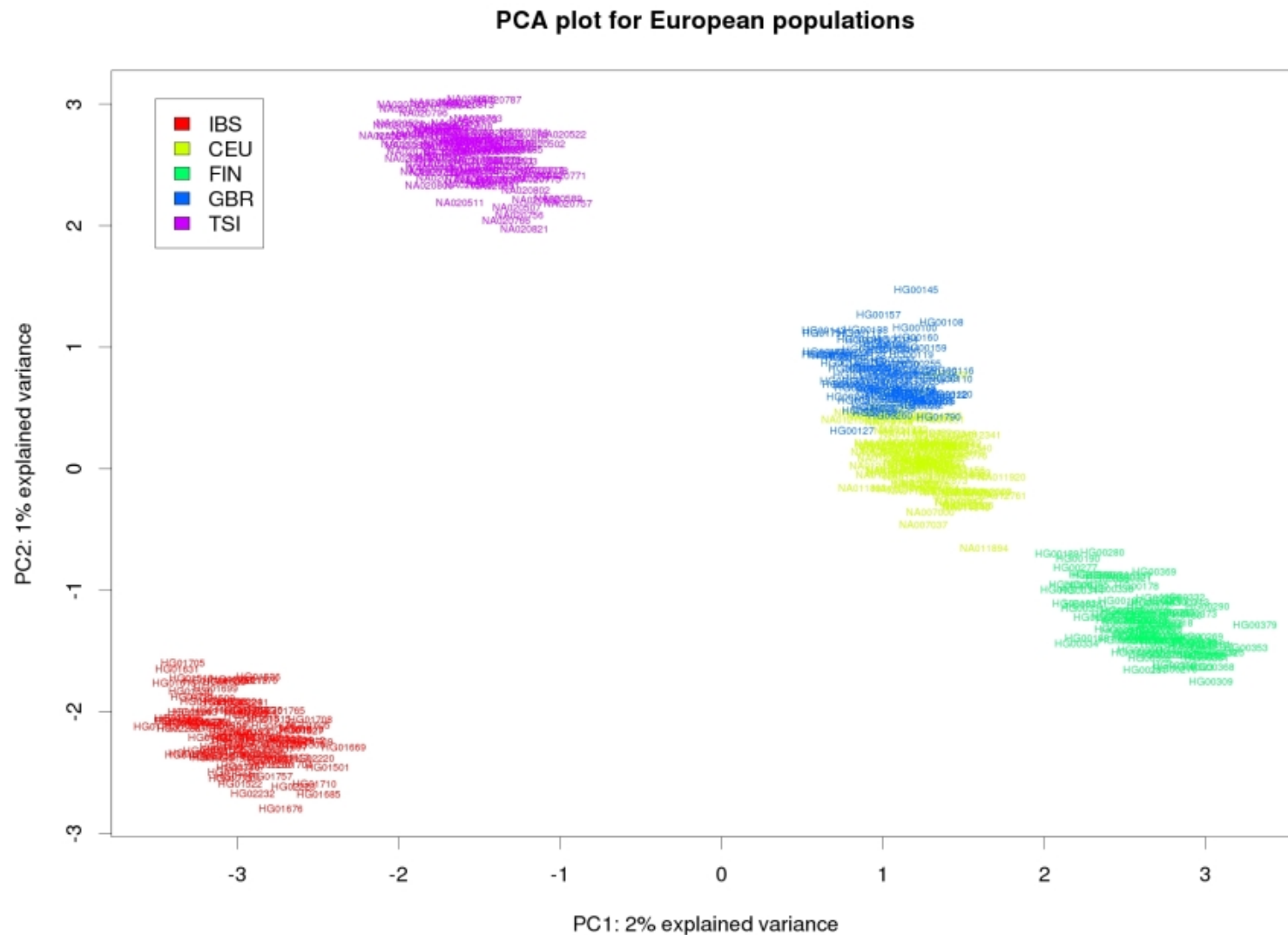
Targeted Enrichment Analysis and Management

CSVS: **CIBERER Spanish Variant Server**

Repositorio de frecuencias de variantes
en la población española

<http://csvs.babelomics.org/>

CIBERER Spanish Variant Server



CSVS

Local genetic variability

Tool interface

Spanish Population Variant Server **beta** Search Studies Stats ?

CLEAR SEARCH

Position

Chromosomal Location:
1:1-100000

Gene:
BRCA2, PPL

Search gene Q

Studies

- ☒ Mgp
- ☒ Virginia Nunes
- ☒ Miguel Angel Moreno
- ☒ Aurora Pujol
- ☒ Francesc Palau

Diseases

☐ Healthy Population

Chr	Position	Alleles	Id	MAF	1000G						EVS					
					Genotypes			Freq.			Genotypes			Freq.		
					0/0	0/1	1/1	0 freq	1 freq	MAF	0/0	0/1	1/1	0 freq	1 freq	MAF
1	17483	C>T		403	1	.	0.917	0.083	0.083							
1	18422	T>C		397	6	1	0.733	0.267	0.267							
1	18256	T>G		403	1	.	0.633	0.033	0.033							
1	18256	T>C		394	10	.	0.633	0.333	0.333							
1	18094	C>T		401	3	.	0.900	0.100	0.100							
1	17398	C>A		399	5	.	0.833	0.167	0.167							
1	16974	C>T		394	10	.	0.667	0.333	0.333							
1	16809	C>G		393	9	2	0.567	0.433	0.433							
1	16794	G>A		403	1	.	0.967	0.033	0.033							
1	16619	C>T		402	.	2	0.867	0.133	0.133							

Genomic Context Effect Frequencies Phenotype

Gene Name	Ensembl Gene Id	Ensembl Transcript Id	Conseq. type	Relative Position	Codon	Strand
Page 0 of 1						

Variants per Study

0k 200k 400k 600k 800k

Variants

<http://csvs.babelomics.org/>

CSVS

CIBERER Spanish Variant Server

Hands on

<http://csvs.babelomics.org/>

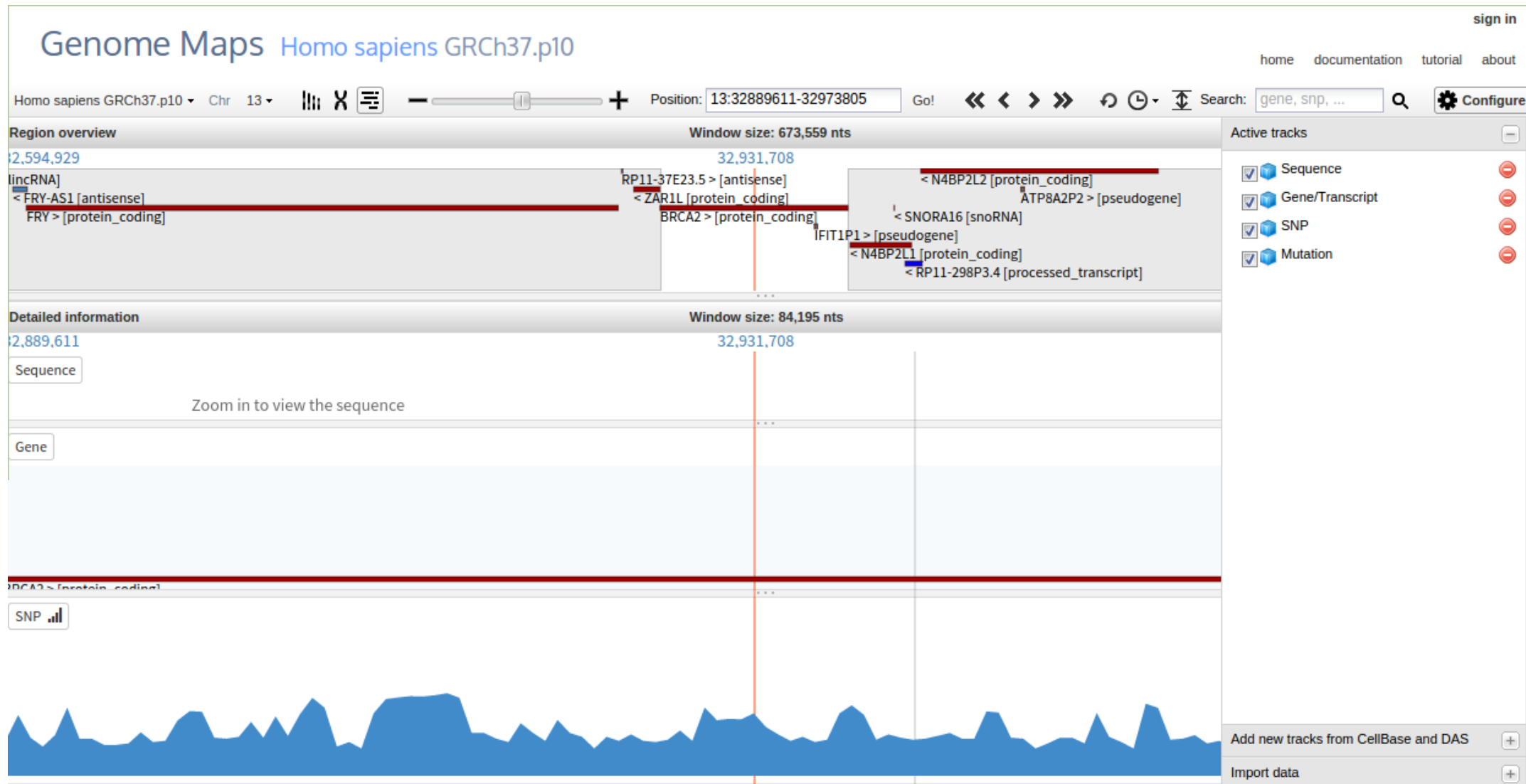
- 1) How many variants do you find in region: 1:24400-70000? (33 variants)
- 2) What information does CSVS give us for this position 1:24536? (Effect, phenotype...)

Genome Maps

Visualizador genómico que interactúa
con bases de datos funcionales

<http://genomemaps.org/>

Tool interface



Genome Maps

A next-generation web-based genome browser

Hands on

<http://genomemaps.org/>

- 1) Visualize this region: 1:100000-200000
- 2) Visualize this gene: LIN28A
- 3) Add new tracks: miRNA, TFBS

Cell Maps

Herramienta de modelización y
visualización de redes biológicas

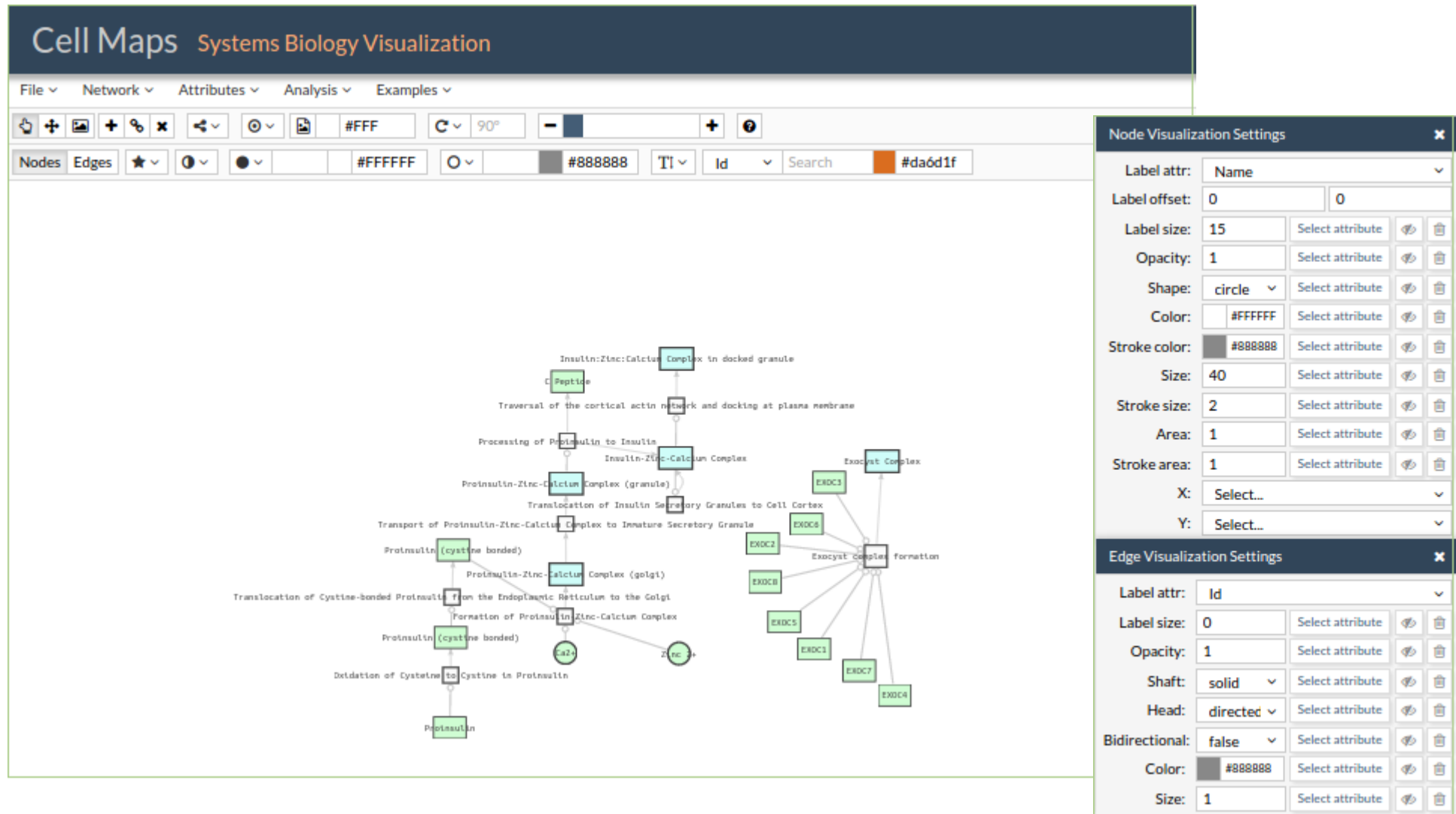
<http://cellmaps.babelomics.org/>

Cell Maps

- 1) Es una herramienta que permite la integración, visualización y el análisis de redes biológicas.
- 2) El **input** es un fichero donde indicamos las relaciones entre los nodos de nuestra red. Opcionalmente podemos incluir un fichero con los atributos de cada nodo.
- 3) El **output gráfico** es una red en la que se muestran las relaciones de los distintos nodos que la integran.

Tutorial: <https://github.com/openCB/cell-maps/wiki>

Tool interface



Cell Maps

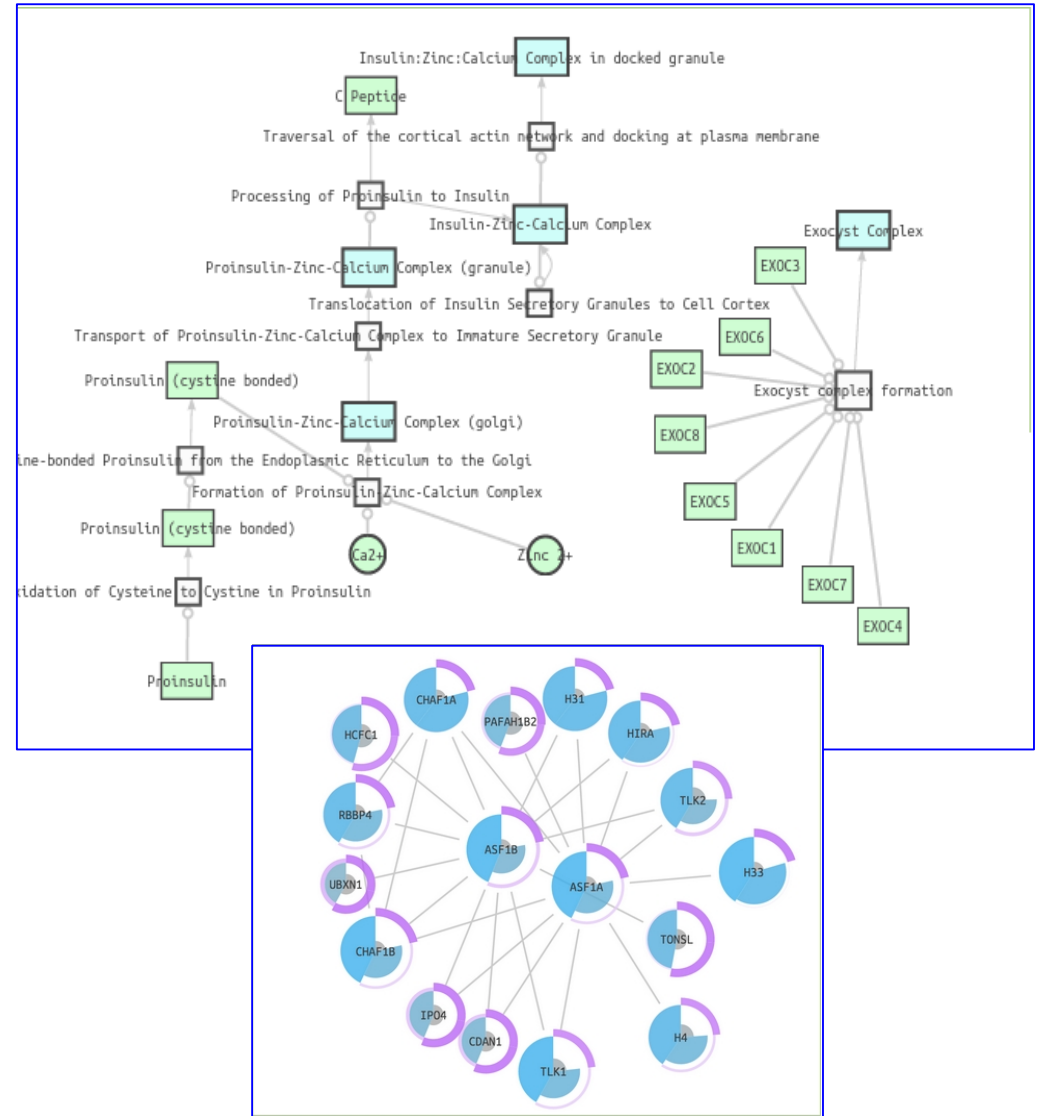
Visualizing and integrating biological networks

Cell Maps: inputs

GO:0000001» pp» GO:0003674
GO:0000001» pp» GO:0005575
GO:0000001» pp» GO:0008150
GO:0003674» pp» GO:0004871
GO:0004871» pp» GO:0038023
GO:0038023» pp» GO:0004888
GO:0004888» pp» GO:0004930
GO:0003674» pp» GO:0097367
GO:0097367» pp»
GO:0005575» pp»
GO:0005575» pp»
GO:0005575» pp»
GO:0005575» pp»
GO:0042995» pp»
GO:0043005» pp»
GO:0042995» pp»
GO:0005575» pp»

ID	<u>pvalor</u>	indi2	descriptor
GO:0031514	0.001	0.16	motile cilium
GO:0000793	0.013	0.129	condensed chromosome
GO:0043025	0.001	0.1	neuronal cell body
GO:0030425	0.003	0.094	dendrite
GO:0044456	0.026	0.086	synapse part
GO:0043005	0.000	0.08	neuron projection
GO:0042995	0.001	0.067	cell projection
GO:0005856	0.044	0.059	<u>cytoskeleton</u>

Cell Maps: outputs



Cell Maps

Visualizing and integrating biological networks

Web Tools for Genomic Data Analysis

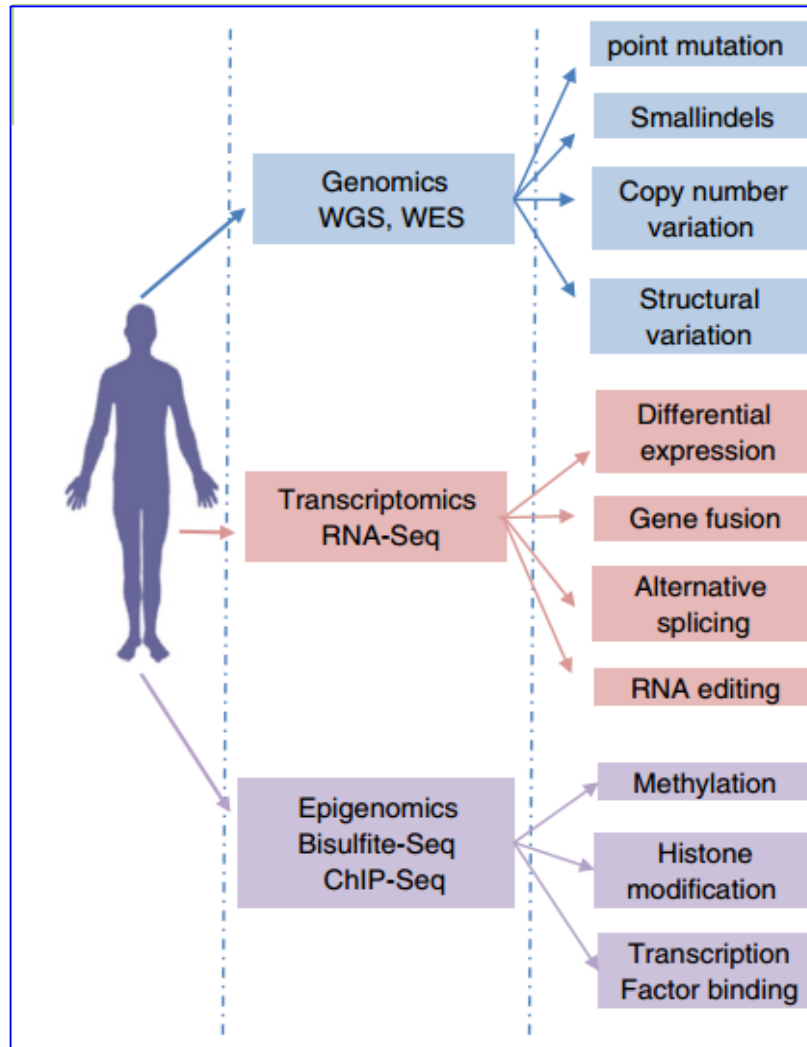
- 1) Introduction to NGS Data Analysis
- 2) RNA-Seq Data Analysis
- 3) Resequencing Data Analysis
- 4) **Omics Data Integration**
 - 1) Ad-hoc approaches
 - 2) Multidimensional Gene Set Analysis
 - 3) Functional Meta-Analysis
 - 4) PATHiVAR
- 5) Functional Profiling

Omics Data Integration from a Systems Biology perspective

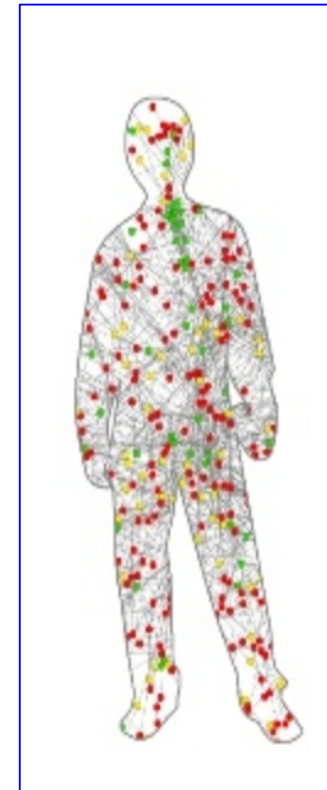


Omic Data Integration

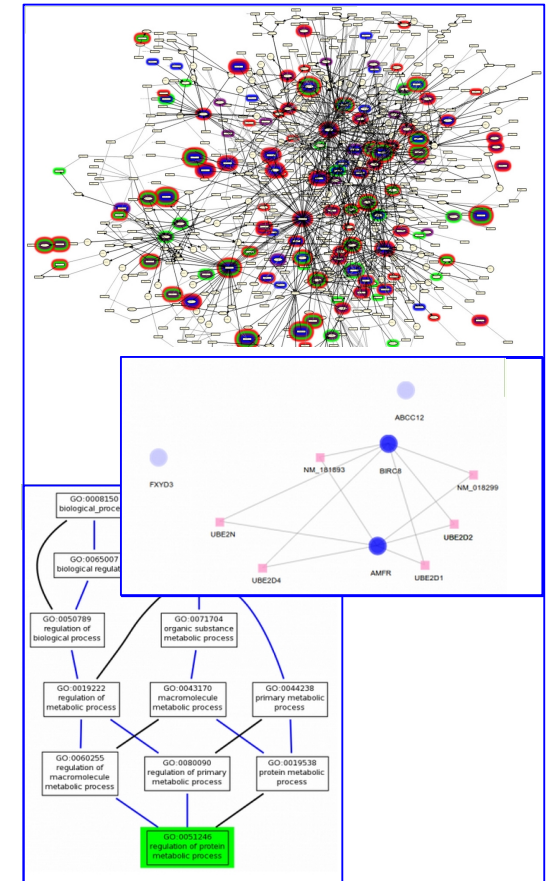
Patient Technologies Data Analysis



Integration and interpretation



Molecular and clinical model



Introduction

Omic Data Integration

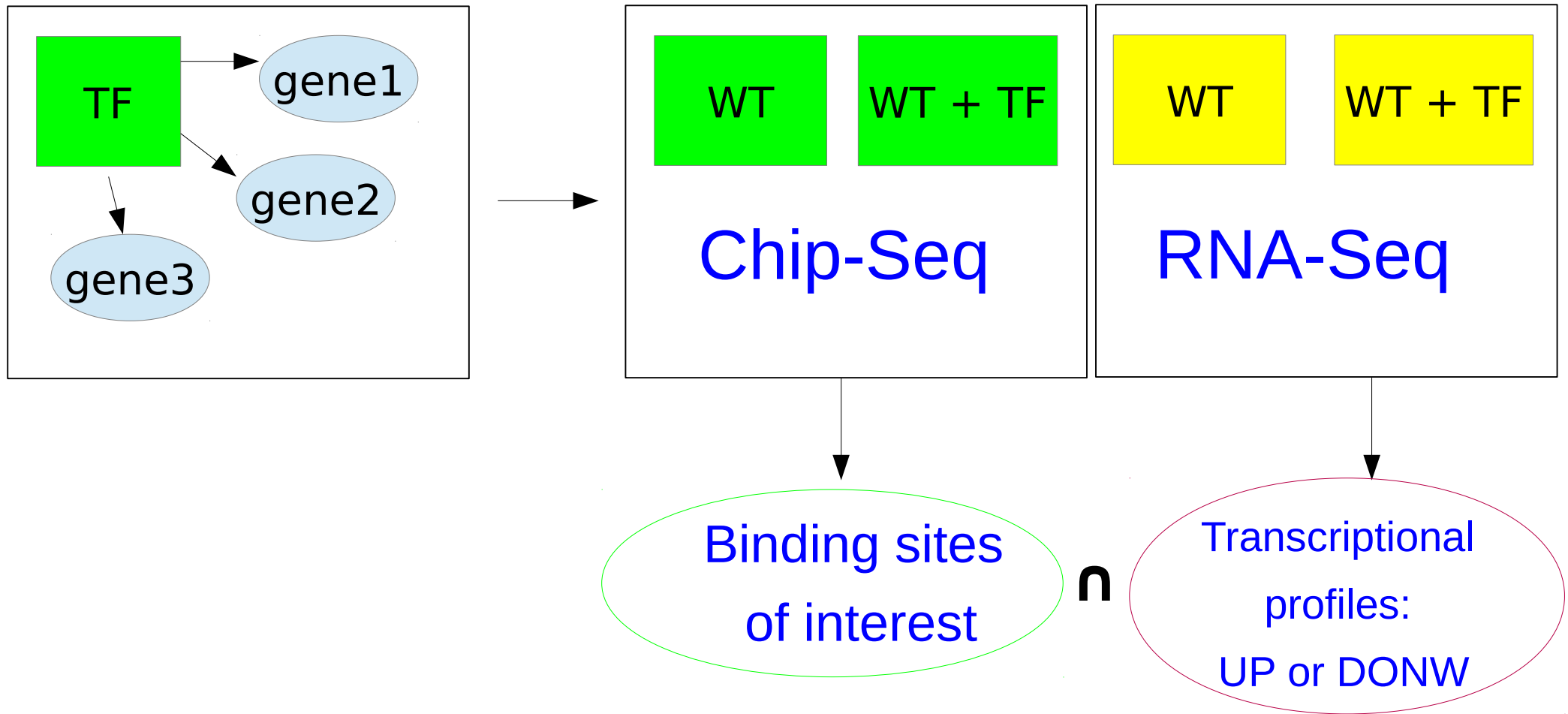
Omics Data Integration

Different strategies:

- 1) Ad-hoc approaches
- 2) Multidimensional Gene Set Analysis
- 3) Functional Meta-Analysis
- 4) PATHiVAR: a web tool to integrate transcriptomics and genomics results

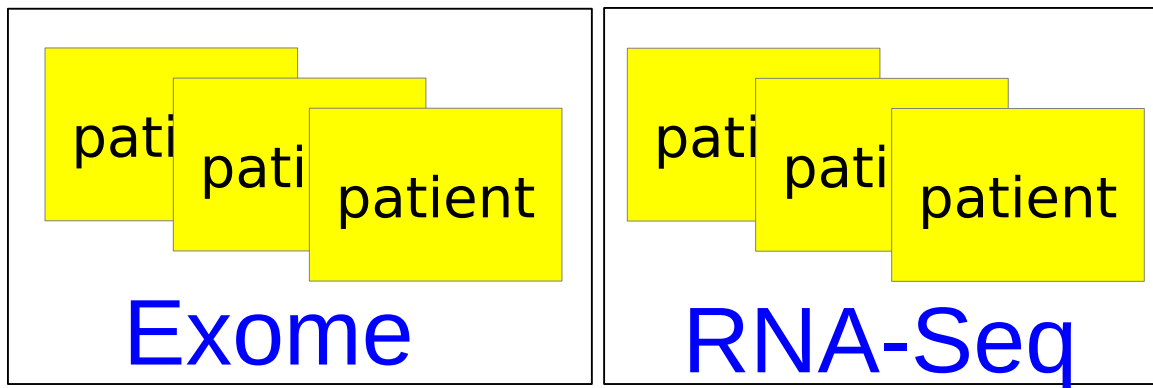
Ad-hoc approaches (1)

Chip-Seq & RNA-Seq



Ad-hoc approaches (2)

Exome & RNA-Seq

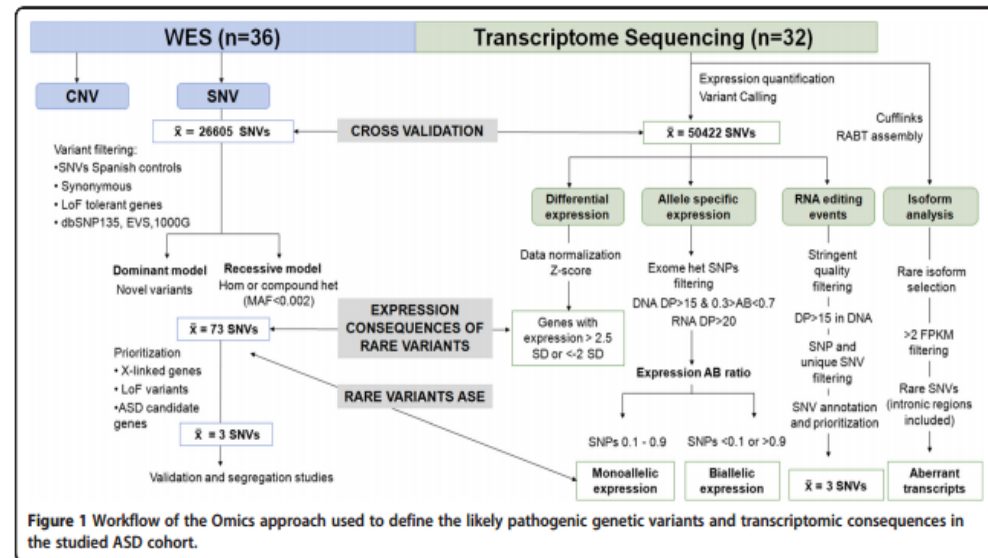


→ Intronic causative mutations

Exonic causative mutations

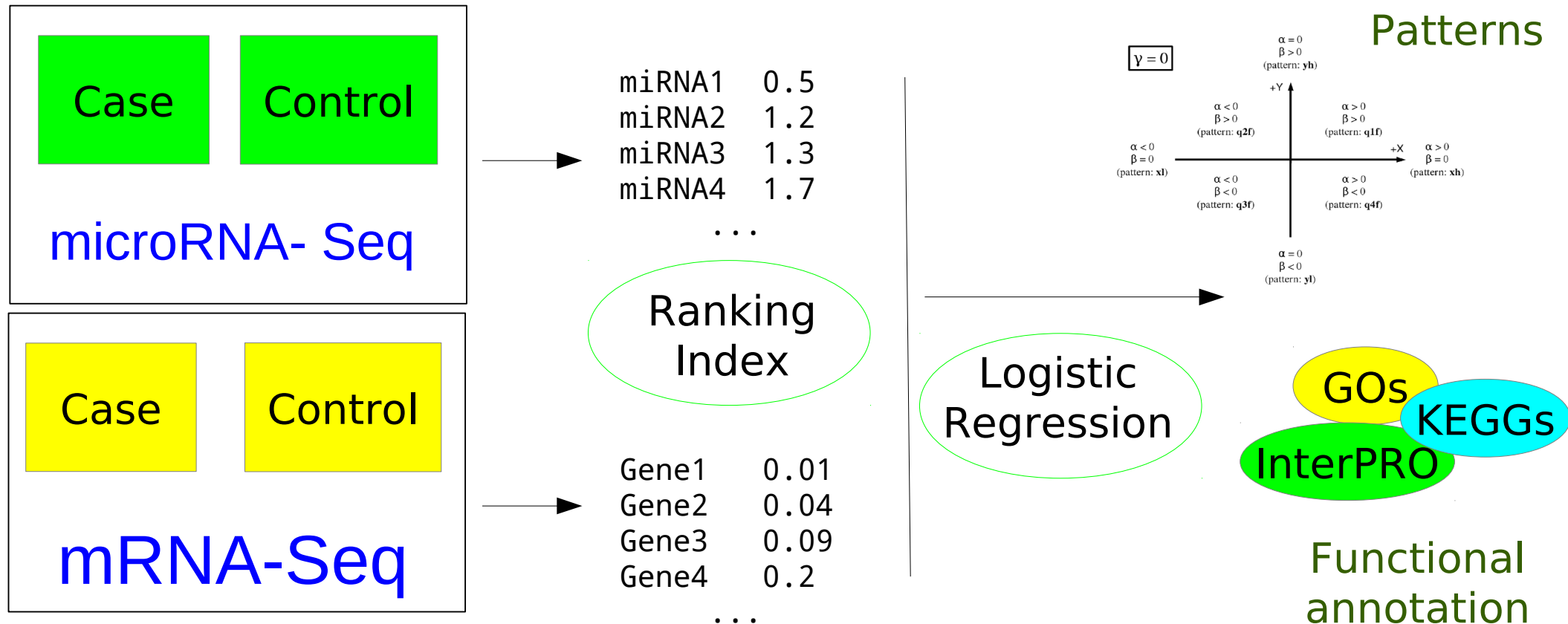
Integrated analysis of whole-exome sequencing and transcriptome profiling in males with autism spectrum disorders

Marta Codina-Solà^{1,2,3}, Benjamín Rodríguez-Santiago⁴, Aïda Homs^{1,2,3}, Javier Santoyo⁵, Maria Rigau¹, Gemma Aznar-Lain⁶, Miguel del Campo^{1,3,7}, Blanca Gener⁸, Elisabeth Gabau⁹, María Pilar Botella¹⁰, Armand Gutiérrez-Arumi^{1,2,3}, Guillermo Antifol^{1,3,5}, Luis Alberto Pérez-Jurado^{1,2,3}* and Ivon Cusó^{1,2,3}*



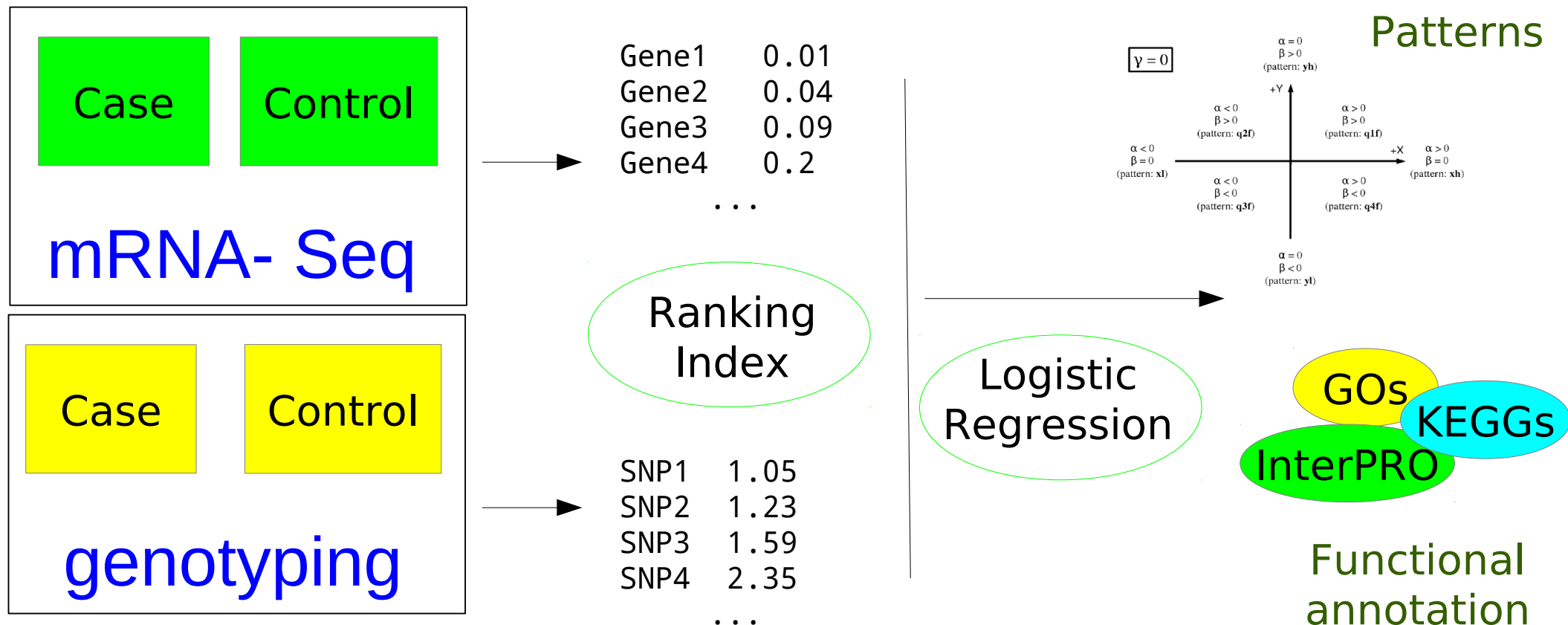
Multidimensional Gene Set Analysis

MicroRNA-Seq & mRNA-Seq



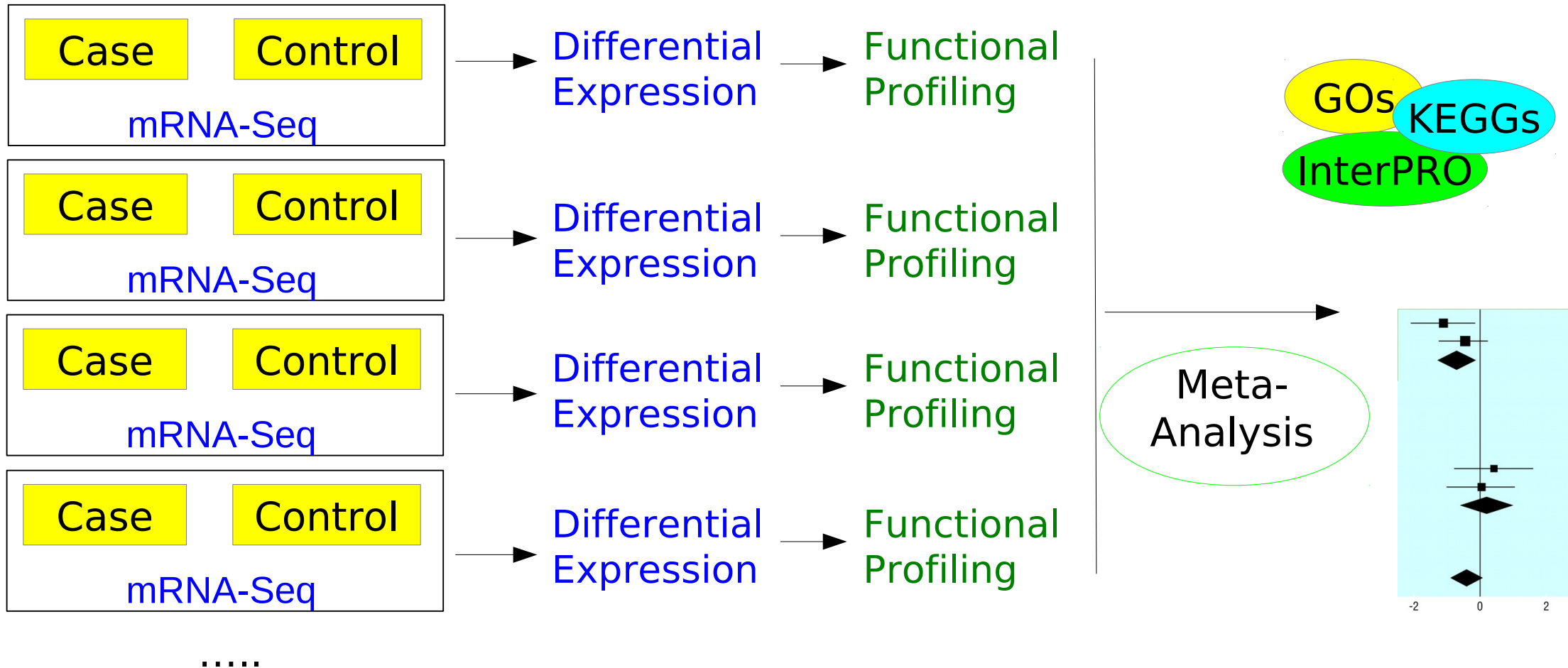
Multidimensional Gene Set Analysis

mRNA-Seq & genotyping association



Functional Meta-Analysis

N mRNA-Seq studies



Strategies

Omics Data Integration

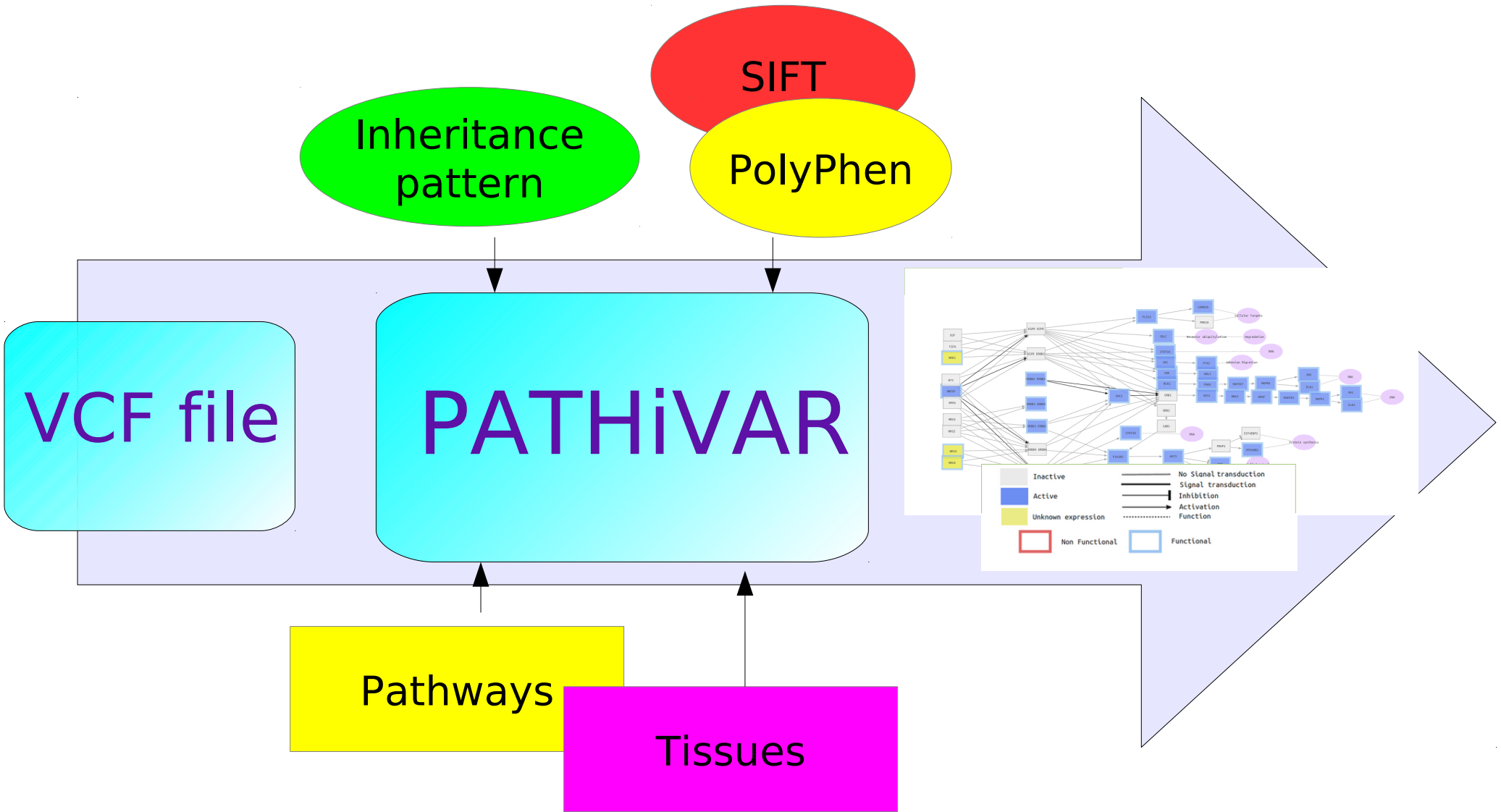
PATHiVAR: mutations and expression

- **PATHiVAR** estimates the functional impact that mutations have over the human signalling network.

- **PATHiVAR:**
 - ▢ Analyses VCF files
 - ▢ Extract the deleterious mutations
 - ▢ Locate them over the signalling pathways in the selected tissue (with the appropriate expression pattern)
 - ▢ Provide a comprehensive, graphic and interactive view of the predicted signal transduction probabilities across the different signalling pathways.

<http://pathivar.babelomics.org/>

How does PATHiVARK work?

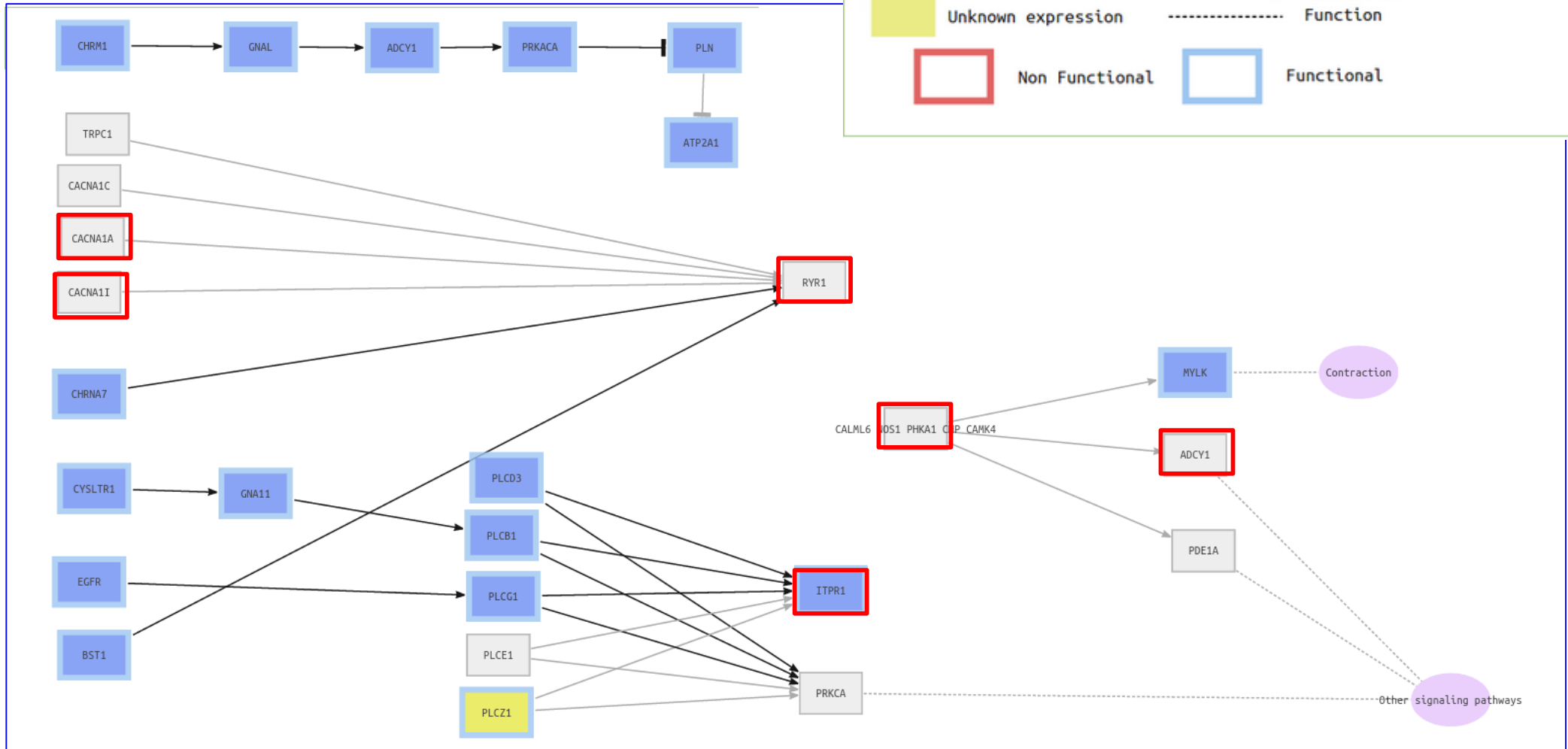


Strategies

PATHiVAR

PATHiVAR

CALCIUM SIGNALING PATHWAY



Strategies

PATHiVAR

More information

OPEN ACCESS Freely available online



Multidimensional Gene Set Analysis of Genomic Data

David Montaner^{1,2}, Joaquín Dopazo¹

Nucleic Acids Research Advance Access published April 16, 2015

Nucleic Acids Research, 2015 **1**
doi: 10.1093/nar/gkv349

Assessing the impact of mutations found in next generation sequencing data over human signaling pathways

Rosa D. Hernansaiz-Ballesteros¹, Francisco Salavert^{1,2}, Patricia Sebastián-León¹, Alejandro Alemán^{1,2}, Ignacio Medina³ and Joaquín Dopazo^{1,2,4,*}

PATHiVAR tutorial:

<http://pathivar.babelomics.org/>



Strategies

Omics Data Integration

Web Tools for Genomic Data Analysis

- 1) Introduction to NGS Data Analysis
- 2) RNA-Seq Data Analysis
- 3) Resequencing Data Analysis
- 4) Omics Data Integration
- 5) Functional Profiling**

Functional Profiling from Babelomics (I)

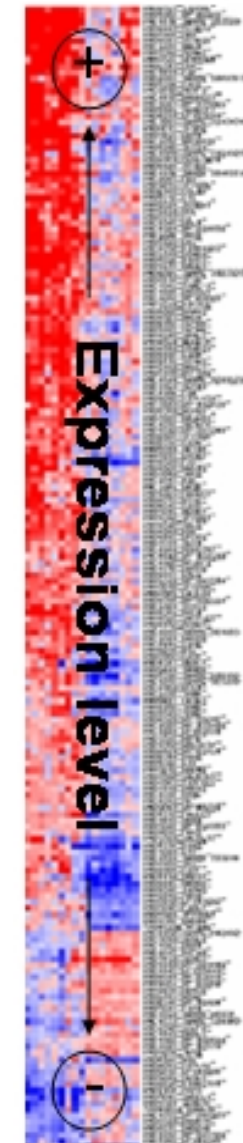
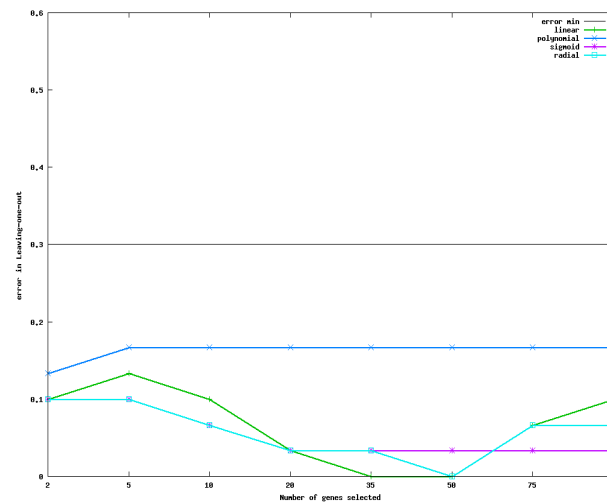
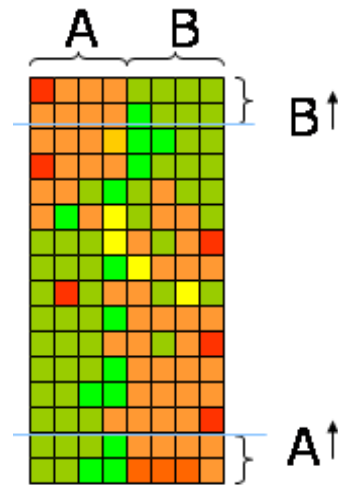
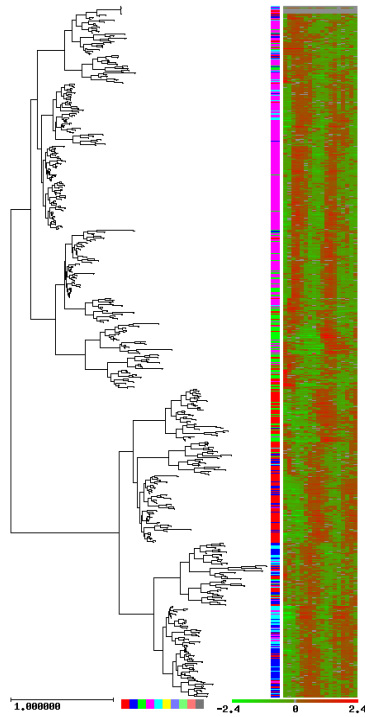


PRINCIPE FELIPE
CENTRO DE INVESTIGACION

Computational • Genomics



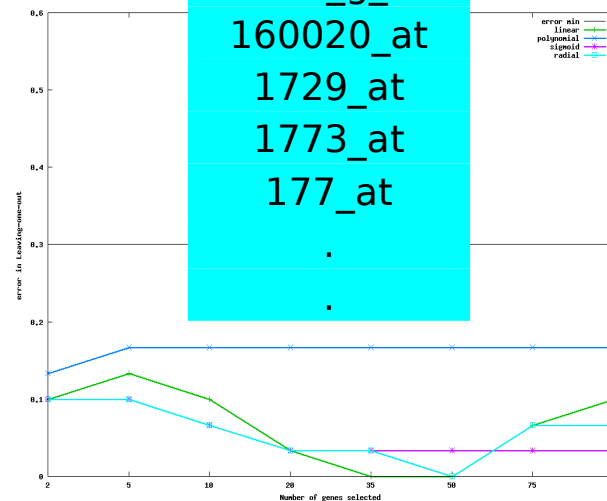
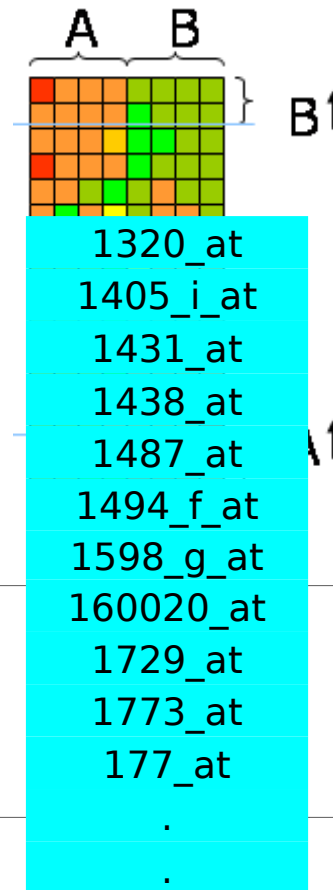
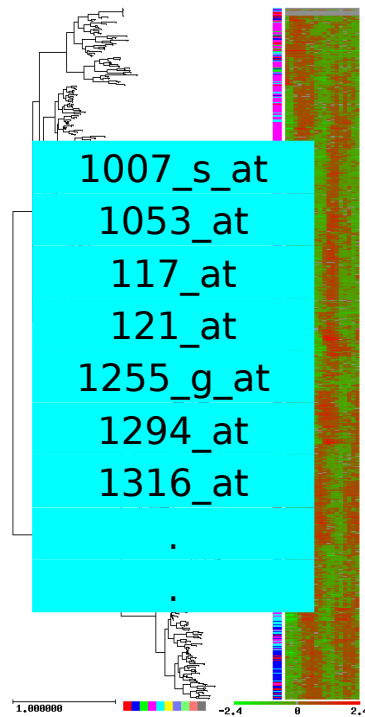
Genome-scale experiment output



Introduction

Functional Profiling

Genome-scale experiment output



1007_s_at	12.4
1053_at	11.5
117_at	10.3
121_at	10.2
1255_g_at	9.9
1294_at	9.3
1316_at	8.2
1320_at	8.1
1405_i_at	7.7
1431_at	7.4
1438_at	6.5
1487_at	6.2
1494_f_at	5.9
1598_g_at	5.8
160020_at	4.8
1729_at	4.7
.	.
.	.

Introduction

Functional Profiling

Functional databases



UniProt/Swiss-Prot
 UniProtKB/TrEMBL
 Ensembl IDs

EntrezGene
 Affymetrix
 Agilent

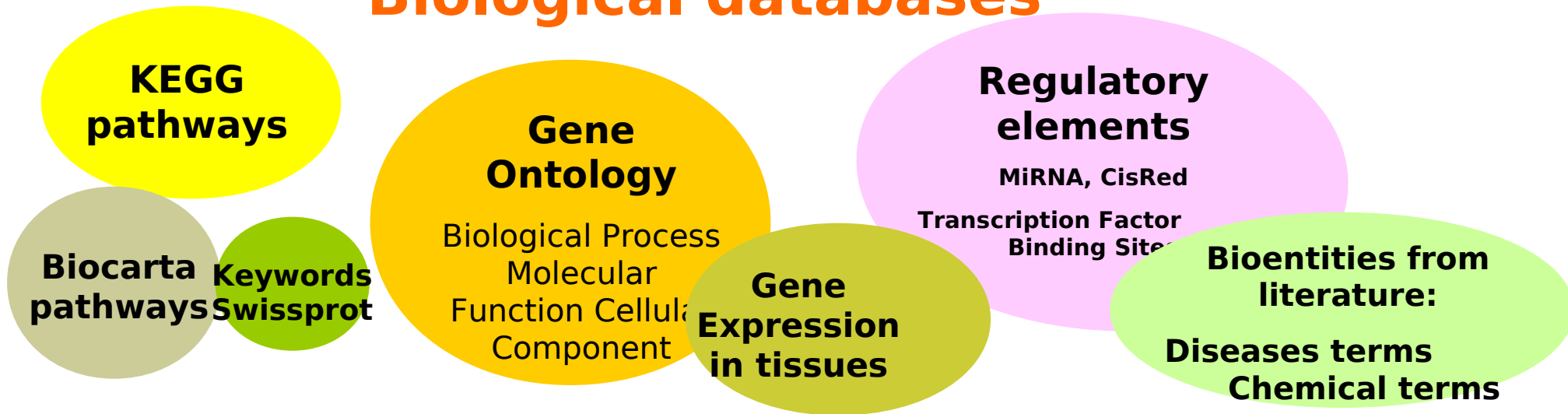


Genes
IDs

HGNC symbol
 EMBL acc
 RefSeq

PDB
 Protein Id
 IPI....

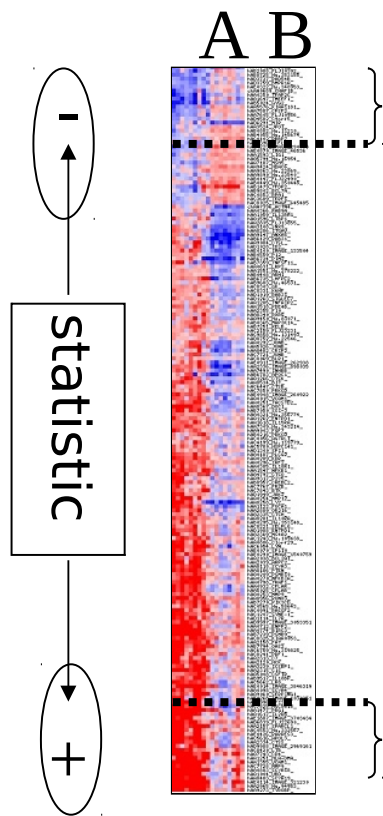
Biological databases



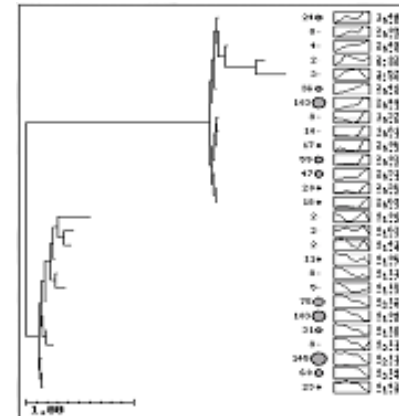
Introduction

Functional Profiling

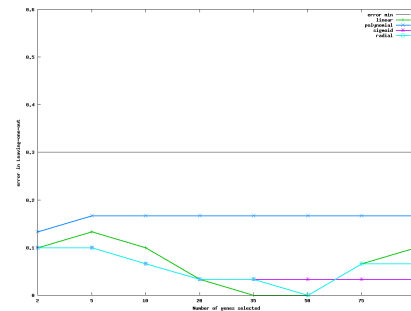
Over-representation analysis



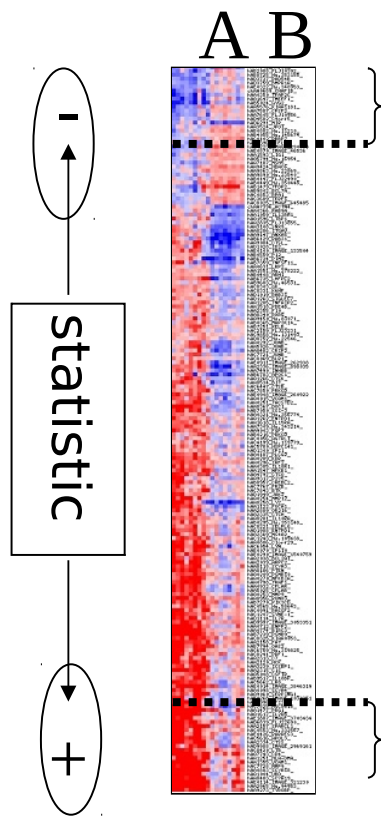
1007_s_at
1053_at
117_at
121_at
1255_g_at
1294_at
1316_at
.
.



1320_at
1405_i_at
1431_at
1438_at
1487_at
1494_f_at
1598_g_at
160020_at
1729_at
1773_at
177_at
.
.

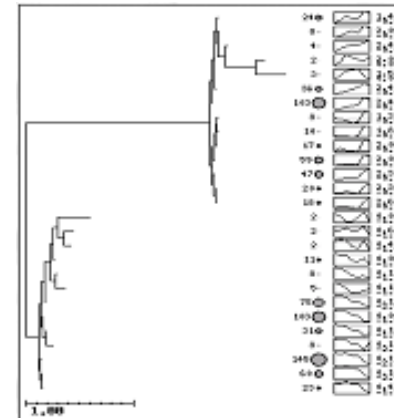


Over-representation analysis

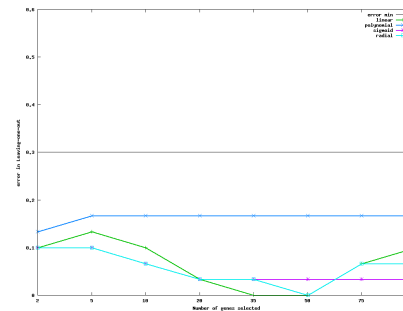


1007_s_at
1053_at
117_at
121_at
1255_g_at
1294_at
1316_at
.
.

Function
4/7



1320_at
1405_i_at
1431_at
1438_at
1487_at
1494_f_at
1598_g_at
160020_at
1729_at
1773_at
177_at
.
.

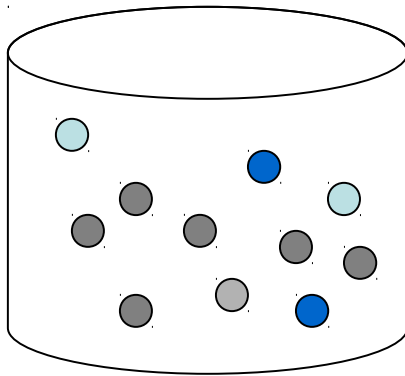


Function
2/11

Over-representation analysis

FatiGO test

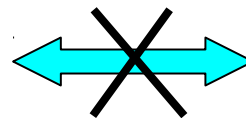
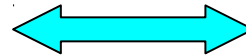
One Gene List (A)



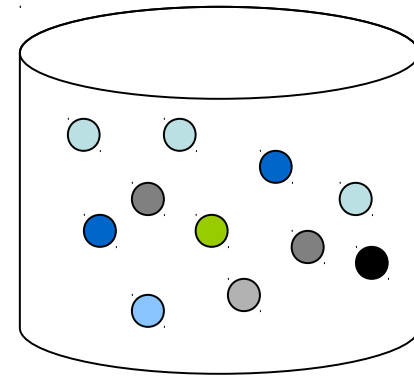
Biosynthesis 60% ●

Sporulation 20% ●

Are this two
groups of genes
carrying out
different
biological roles?



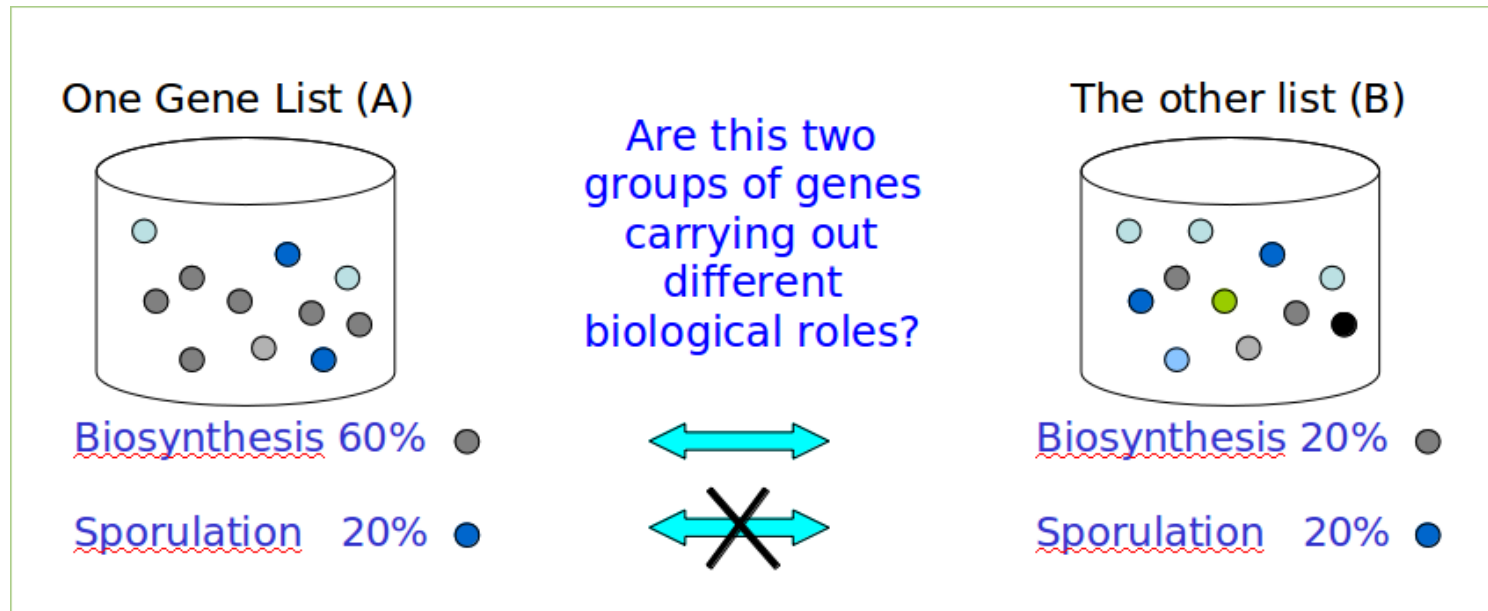
The other list (B)



Biosynthesis 20% ●

Sporulation 20% ●

Over-representation analysis

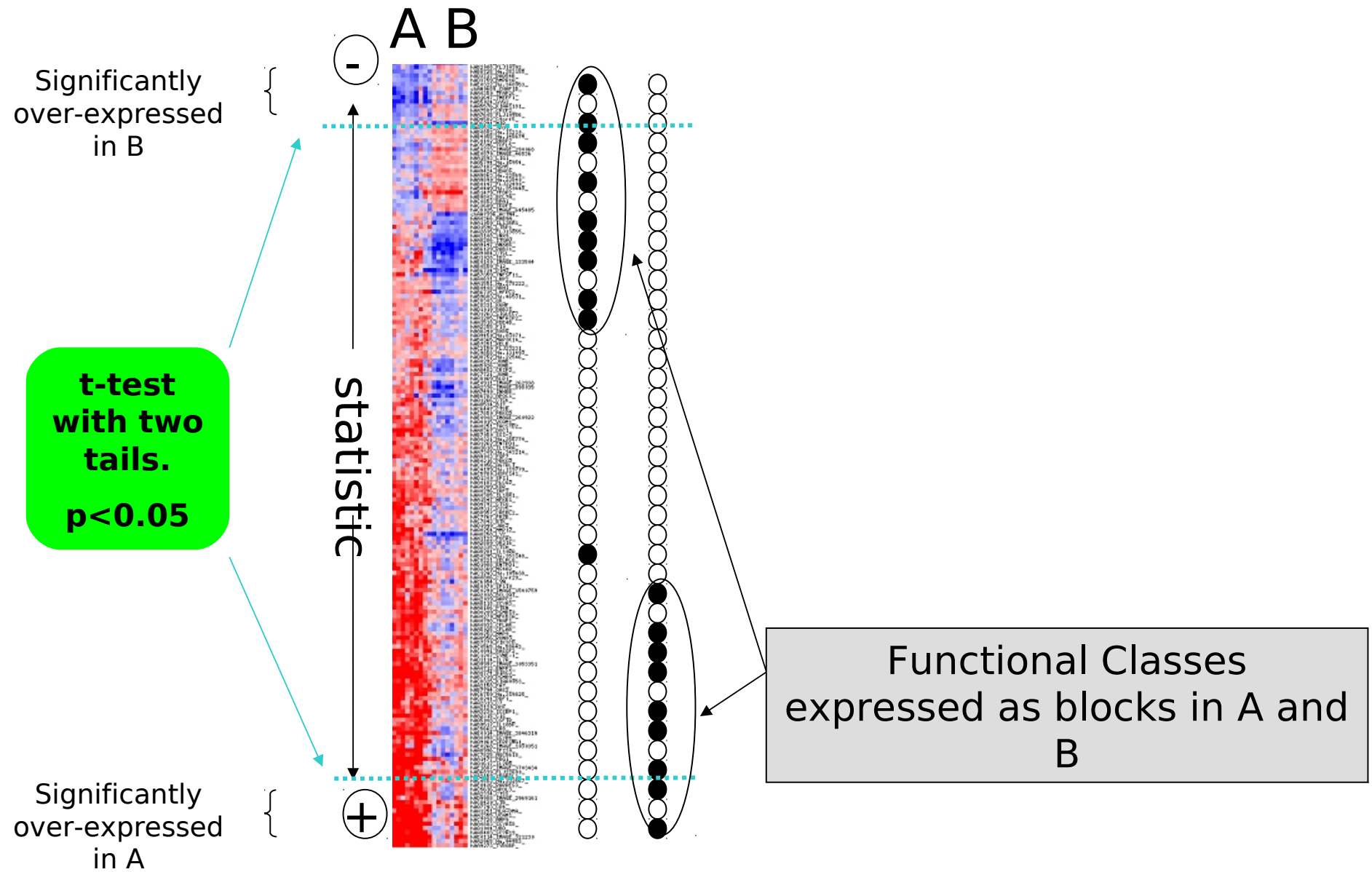


Genes in group A have significantly to do with biosynthesis, but not with sporulation.

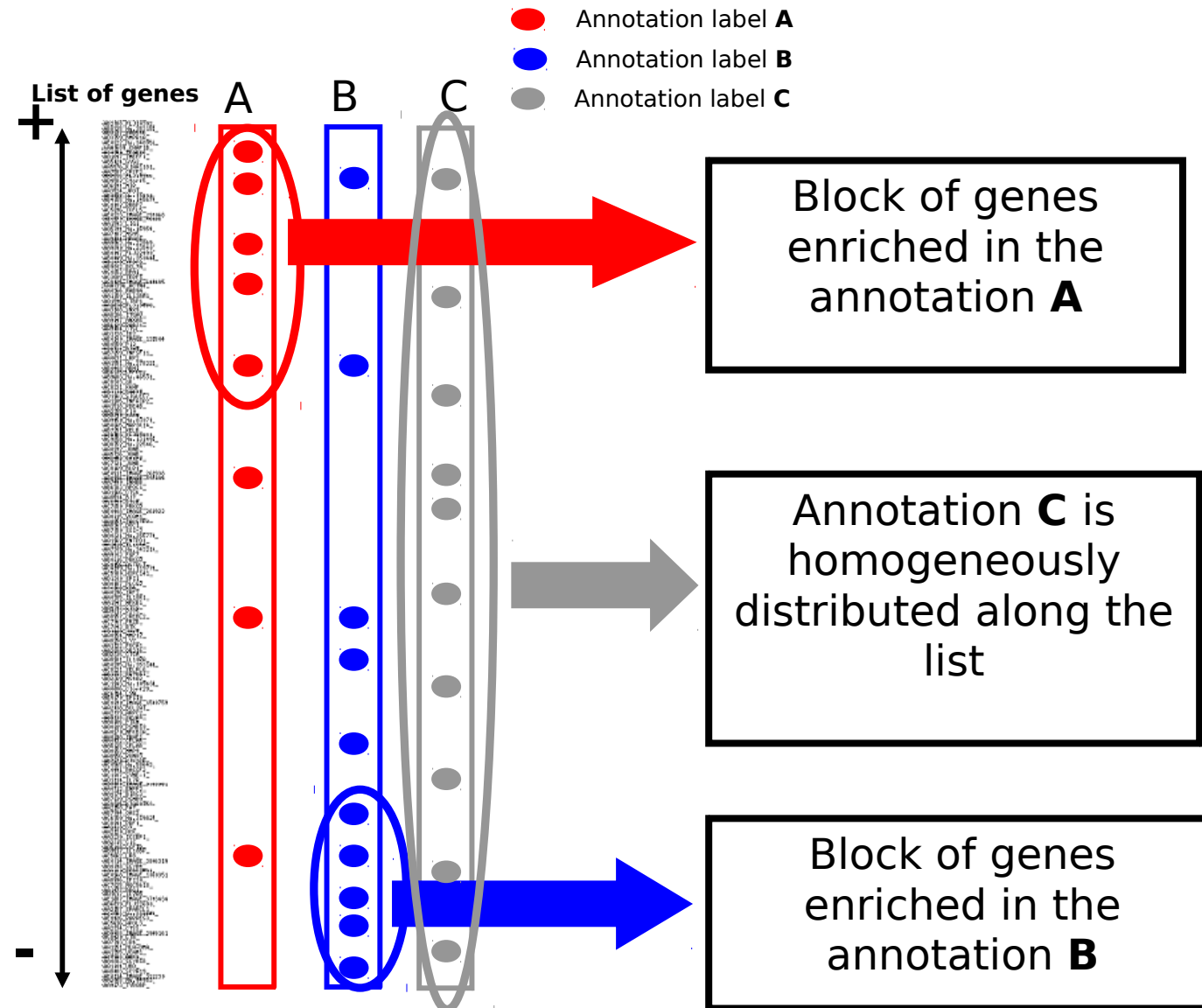
	A	B
Biosynthesis	6	2
No biosynthesis	4	8

We do this for each term (GO, miRNA, Interpro , ...)
Thousand of terms, so Multiple Test Correction is needed!!!

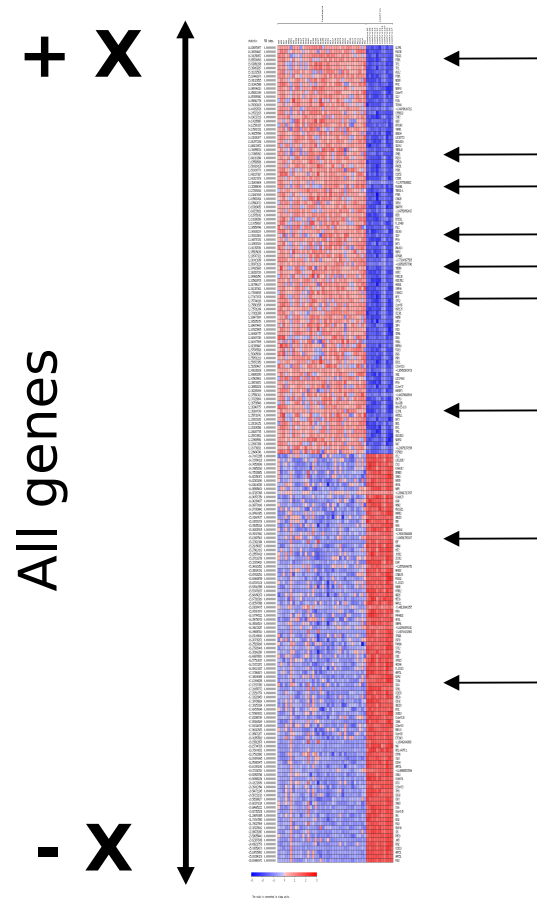
Gene Set Analysis



Gene Set Analysis



Gene Set Analysis



$$\ln \left(\frac{P(g \in F)}{P(g \notin F)} \right) = K + \alpha X$$

alpha > 0 : increasing X increases the probability of the gene to be annotated

alpha < 0 : decreasing X increases the probability of the gene to be annotated

Hands on



Babelomics 5

<http://babelomics.bioinfo.cipf.es/>

Functional / FatiGO
Functional / Logistic Model

Online examples

Babelomics 5

Functional Profiling

Functional Profiling from Babelomics (II)



PRINCIPE FELIPE
CENTRO DE INVESTIGACION

Computational • Genomics



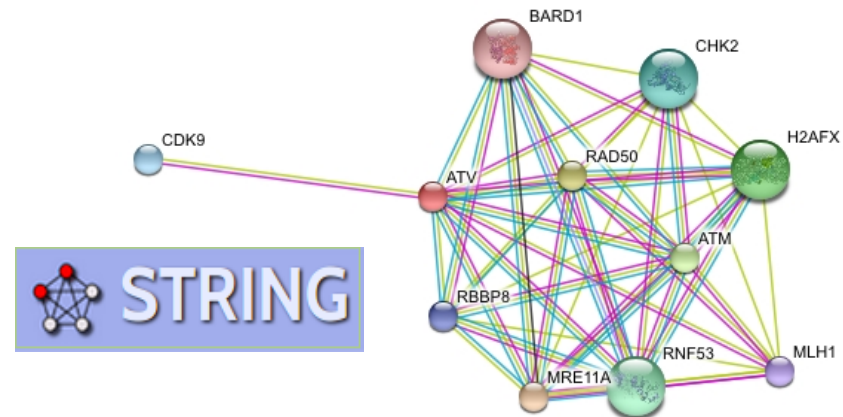
Protein-Protein Interactions (PPI)

- PPIs are a central point at almost every level of cell function:
 - Structure of subcellular organelles (structural proteins)
 - Packing the chromatin (histones)
 - Protein modifications (kinases)
- Retrieving information about a **single protein**....

5/277 Interacting proteins for BRCA1 (ENSP00000350283³)

Interactant		Interaction
GeneCard	External ID(s)	
NBN	ENSP00000265433 ³	STRING (score=.
TOPBP1	ENSP00000260810 ³	STRING (score=.
UBA1	ENSP00000338413 ³	STRING (score=.
UBE2D1	ENSP00000185885 ³	STRING (score=.
GADD45A	ENSP00000360025 ³	STRING (score=.

[About this table](#)

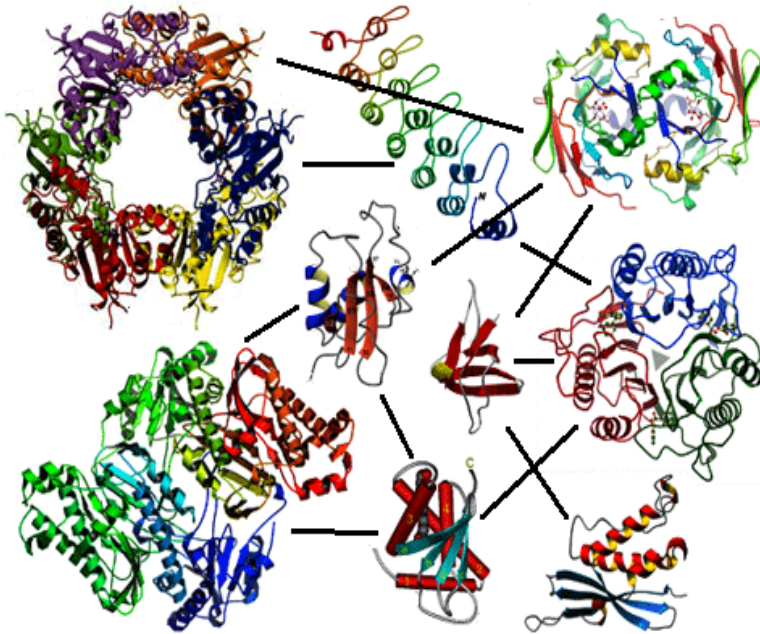


Protein-Protein Interactions (PPI)

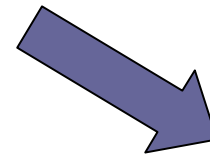
- How to extract information about **sets** of genes?
- How to perform **functional enrichment analysis** using protein-protein interactions as annotation source?
- How to **prioritize candidate genes**?
- How to get **new functional candidate genes**?

Graph Theory

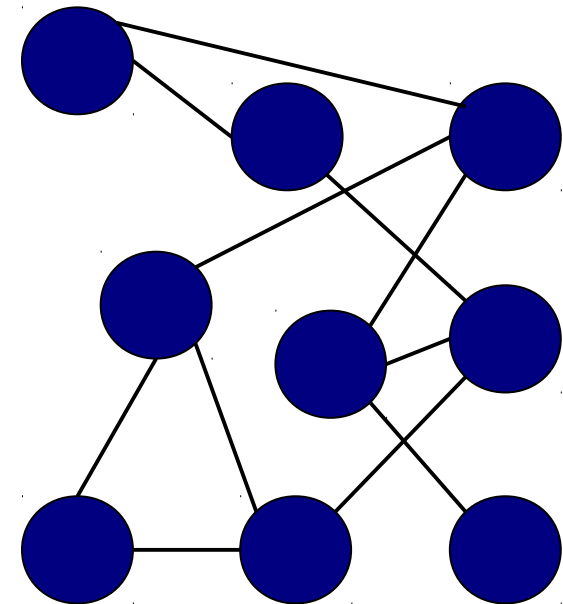
Set of proteins interacting



Nodes = proteins
Edges = interaction events



Undirected graph



structured data

Graph Theory

Graph theory may help us to study protein networks.
Some interesting parameters:

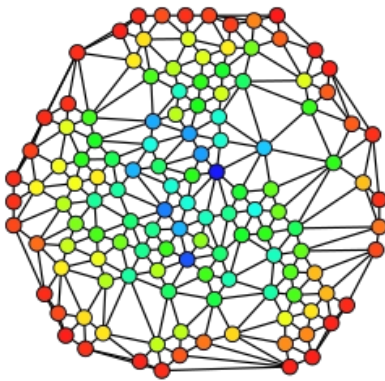
- **Degree (connectivity or connections)**: number of edges connected to a node. Nodes with high degree are called **hubs**.

- **Betweenness**: A measure of centrality of a node, it is defined by:

$$C_B(v) = \sum_{s \neq v \neq t \in V} \frac{\sigma_{st}(v)}{\sigma_{st}}$$

σ_{st} is total number of shortest paths in the graph.

$\sigma_{st}(V)$ is the number of shortest paths that pass through node V

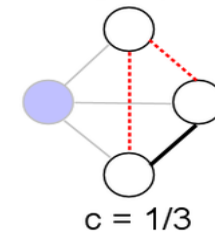


Graph Theory

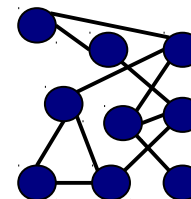
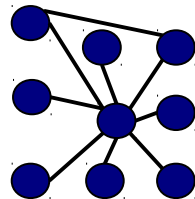
- **Clustering coefficient (of a node)**: A measure of how interconnected the neighbours of that node are. Proportion of links between the nodes within its neighbourhood divided by the number of links that could possibly exist between them.

$$C_i = \frac{2e_i}{n_i(n_i - 1)}$$

e_i is the number of edges among the nodes connected to node i
 n_i is the number of neighbours of node i



To differentiate between **star-shaped** nets and more **interconnected** nets.



Graph Theory

Some Graph Theory concepts:

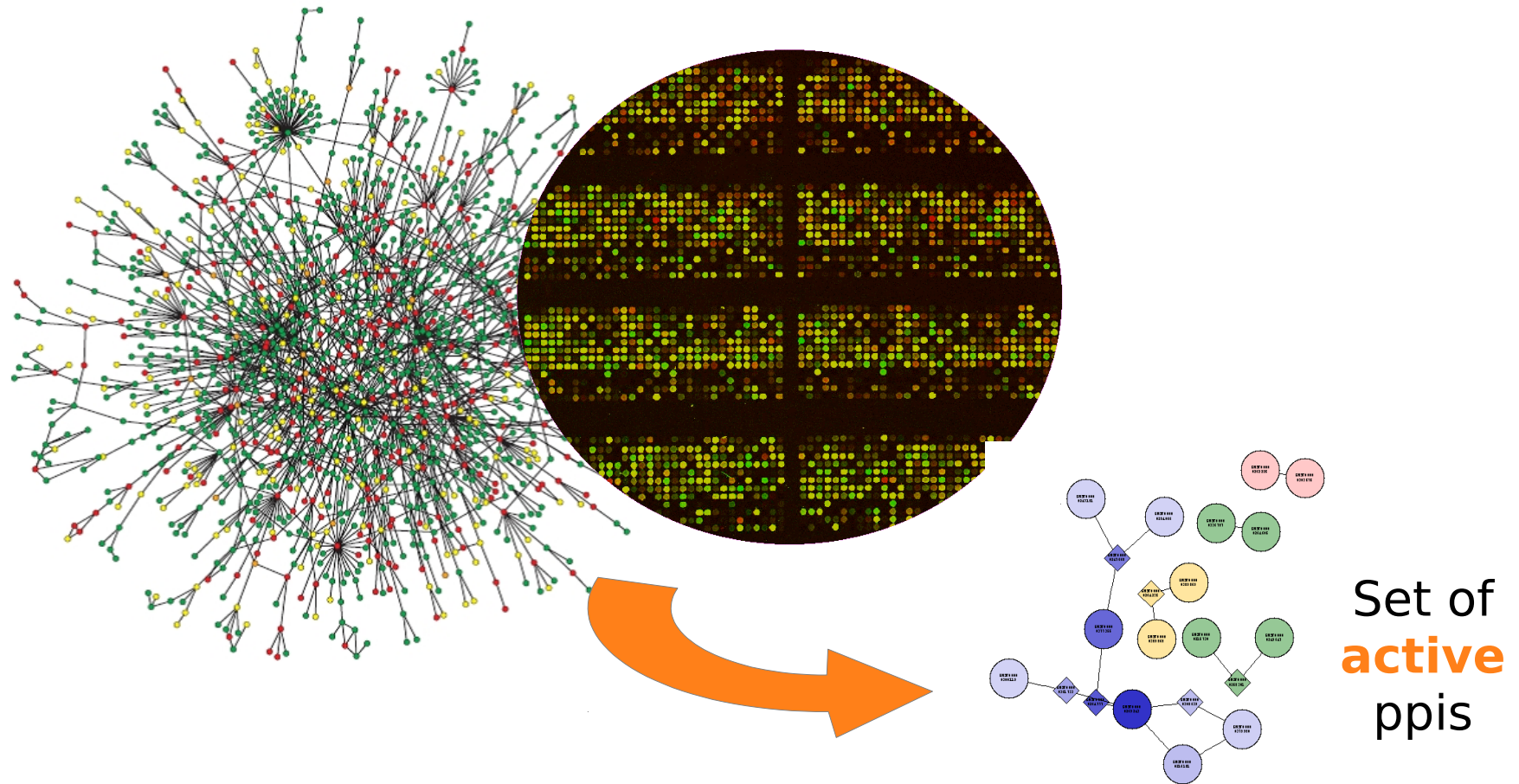
Shortest path. The path with less edges that connects two nodes.

Component. A group of nodes connected among them.

Bicomponent. A group of nodes connected to other group of nodes by only an edge. The edge that joins two bicomponents is called **articulation point**.

Interactome & Transcriptome

- **Interactome.** Complete collection of protein-protein interactions in the cell.
- **Transcriptome** determines the real interactome.



Interactome & Transcriptome

Goal

To develop a methodology that may **extract from lists of proteins/genes** the ppi networks acting and evaluates whether they have importance in the **cooperative behaviour** of the list.

How we evaluate the cooperative behaviour of a list of proteins/genes in terms of its ppi network parameters?

Two different approximations

- Importance in **complete interactome**
- Cooperative behaviour - **Minimal Connected Network**

Network Analysis: SNOW



Babelomics 5

<http://babelomics.bioinfo.cipf.es/>

Functional / Network Enrichment:
SNOW

Hands on

There is a well-known list of 72 genes related to eye diseases (ABCA4, ABHD12, ADAMTS18, AIPL1, BBS1, BEST1, C2orf71, C8ORF37, CA4, CABP4, CEP290, CERKL, CHM,...)

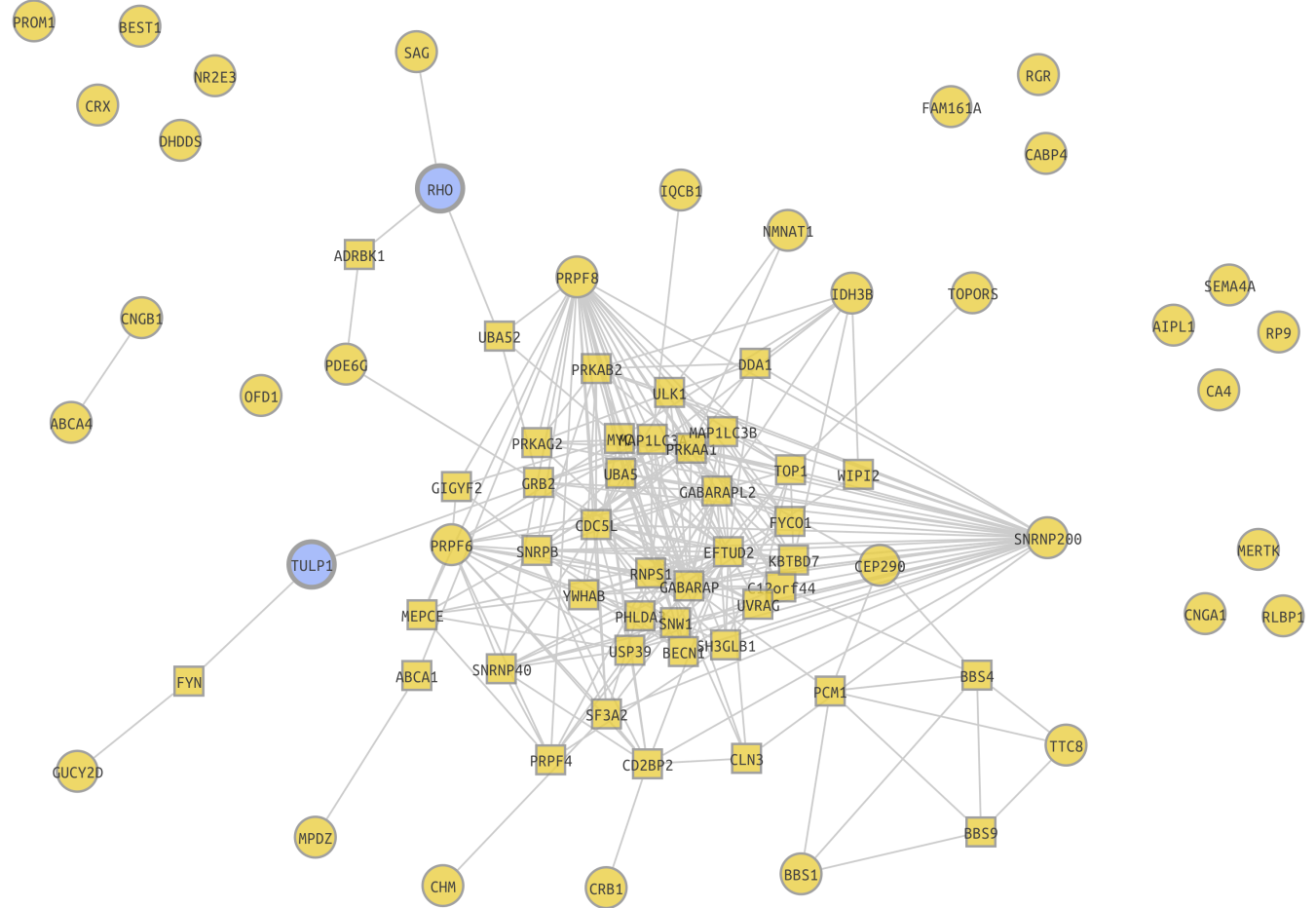
- 1) Now we have a two new candidates: RHO and TULP1 . We would to know what is the relationship between all genes.
- 2) Also it would be interesting to explore new functional candidates.

Strategies from Babelomics?

- Single Enrichment
- **Network** Enrichment

Hands on

RHO	TULP1
ABCA4	MERTK
ABHD12	MPDZ
ADAMTS18	NMNAT1
AIPL1	NR2E3
BBS1	NRL
BEST1	OFD1
C2orf71	PDE6A
C8ORF37	PDE6B
CA4	PDE6G
CABP4	PRCD
CEP290	PROM1
CERKL	PRPF3
CHM	PRPF31
CLRN1	PRPF6
CNGA1	PRPF8
CNGB1	PRPH2
CRB1	RBP3
CRX	RD3
CYP4V2	RDH12
DHDDS	RGR
EYS	RLBP1
FAM161A	ROM1
FSCN2	RP1
GUCA1B	RP2
GUCY2D	RP9
IDH3B	RPE65
IMPDH1	RPGR
IMPG1	RPGRIP1
IMPG2	SAG
IQCB1	SEMA4A
KCNJ13	SNRNP200
KLHL7	SPATA7
LCA5	TOPORS
LRAT	TTC8
MAK	USH2A



More info + questions

Nucleic Acids Research Advance Access published May 26, 2014

Nucleic Acids Research, 2014 **1**
doi: 10.1093/nar/gku472

A web tool for the design and management of panels of genes for targeted enrichment and massive sequencing for clinical applications

Nucleic Acids Research Advance Access published May 6, 2014

Nucleic Acids Research, 2014 **1**
doi: 10.1093/nar/gku407
,^{2,3,*}
, and

A web-based interactive framework to assist in the prioritization of disease candidate genes in whole-exome sequencing studies

Aleja
Joaq

Nucleic Acids Research Advance Access published April 20, 2015

Nucleic Acids Research, 2015 **1**
doi: 10.1093/nar/gkv384

Babelomics 5.0: functional interpretation for new generations of genomic data

¹Comp
²Bioinf
³Func

Published online 8 June 2013

Nucleic Acids Research, 2013, Vol. 41, Web Server issue **W41–W46**
doi:10.1093/nar/gkt530

Genome Maps, a new generation genome browser

Ignac
Robe

Nucleic Acids Research Advance Access published April 16, 2015

Nucleic Acids Research, 2015 **1**
doi: 10.1093/nar/gkv349

Assessing the impact of mutations found in next generation sequencing data over human signaling pathways

OPEN ACCESS Freely available online

PLoS one

Multidimensional Gene Set Analysis of Genomic Data

David Montaner^{1,2}, Joaquín Dopazo^{1,2,3*}



BIER

Tutorial: web tools

Web tools

NGS Data Analysis: RNA, Whole Exome and Gene Panel

Other resources for Genomic Data Analysis

